Author:
Beynon, Rhona A

Title:
Biological and lifestyle predictors of survival in head and neck cancer.

Biological and lifestyle predictors of survival in head and neck cancer.

Rhona Alison Beynon

23rd March 2020

Bristol Medical School: Population Health Sciences.

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Health Sciences.

Word count: 73,390

Abstract

Head and neck cancer (HNC) is a heterogeneous disease with variable clinical outcomes. Several biomarkers for HNC prognosis have been identified and incorporated into clinical prediction models, but the ability of these models to predict outcome could be improved.

Using self-report and peripheral blood-based DNA methylation (DNAm) predictors, I examined whether lifestyle and life-course exposures (smoking, alcohol-drinking, body mass index (BMI), education and biological aging) predict all-cause mortality and should be considered for inclusion in prognostic models. Additionally, I examined the association of circulating metabolites with all-cause mortality.

I found that current smokers with HNC were twice as likely to die during follow-up than never-smokers (HR=2.0; 95% CI: 1.4, 3.0), after controlling for established prognostic predictors. In people with oropharyngeal cancer (OPC), DNAm-based predictors of smoking were associated with all-cause mortality. Based on a DNAm predictor that included 2,623 CpG sites (Joehanes-Bonferroni) and on methylation levels at a single site (cg05575921) within the aryl-hydrocarbon receptor repressor (AHRR) gene, the change in risk of death (all-cause) per unit increase in standardised DNAm score was approximately two-fold, where 1 unit increase corresponds to the average increase in the measured methylation score experienced by current smokers compared with never smokers (HR=1.89 [95% CI:1.06, 3.47] and 1.92 [95% CI:1.03, 2.33]).

Two DNAm-age measures, comprising a set of 1,030 CpGs (*AgeAccelGrim*) and 71 CpGs (*IEAAHannum*), were associated with all-cause mortality in OPC. *AgeAccelGrim* had the largest magnitude of effect in fully adjusted models: each SD increase in *AgeAccelGrim* resulted in a 39% increased all-cause mortality risk (HR=1.39; 95% CI:1.06, 1.83). The addition of *AgeAccelGrim* to a standard clinical model slightly improved mortality risk prediction at 3-years (AUC: 0.80 vs. 0.77; *p*-value for difference=0.069).

In adjusted models, acetate and creatinine were associated with mortality in OPC (HRs per SD increase in metabolic trait =1.30 [95% CI:1.11, 1.51] and 0.68 [95% CI: 0.53, 0.89], respectively).

This thesis demonstrates the potential of lifestyle, epigenetic and metabolomic measures to enhance survival prediction in people with HNC, though these findings need to be replicated in independent clinical cohorts.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Rhona Beynon DATE: 23rd March 2020

Acknowledgements

I am hugely grateful to my supervisors Professor Andy Ness and Professor Richard Martin. To Andy, thank you for giving me the opportunity to work as an intern at the former NIHR Biomedical Research Unit (BRU) in Nutrition, Diet and Lifestyle. Working at the BRU instilled in me the drive and confidence I needed to undertake a research project of my own. Without this experience, I would almost certainly not have been able to apply for an Epidemiology PhD, having no previous experience in the field; though I'm sure my knowledge of "Developmental Biology", "host-parasite interactions" and "life in extreme environments" helped somewhere along the way. To Richard, I thank you for your continued enthusiasm and encouragement, which helped foster a stimulating and enjoyable working environment. You have never doubted my ability, even in the early days when I was having a meltdown over my first mini-project and my aversion to statistics.

On that note, I am grateful to Professor Margaret May, who supervised me for a large part of this PhD and Dr Suzanne Ingle for stepping in when Margaret "went off Piste" to enjoy her retirement. It was not an easy task coming on board so late in my PhD, when I had multiple analyses going on and a million different questions to ask. Margaret, whilst my relationship with statistics is still very much a work in progress, thanks to your influence, I now realise that it is an ally and not an enemy. In fact, I *almost* enjoy statistics, which I never thought I would say, and that is credit to your wonderful teaching and humility. Four years ago, if someone had told me I would be doing this thing called flexible parametric survival modelling, I would have said you'd got the wrong student but thanks to yours and Sue's patience and approachability, I'm doing things I never imagined I'd be capable of and for that I am forever grateful.

I would like to thank Dr Diana Dos Santos Ferreira, both for sharing her vast knowledge of metabolomics and for being a friend. I thoroughly enjoyed working with you and Dr Rebecca Richmond on my lycopene and green tea mini-project. I learnt so much from you both. I am sure I had the honour of working with future Professors!

A special thanks goes to the participants of the Head and Neck 5000 study and the Head and Neck 5000 study team, especially Katrina Hurley, Stu Toms and Kate Ingarfield, who made this research possible. Similarly, the Wellcome Trust for funding my PhD studentship.

Outside of academia, I have so many people to be grateful for: my parents who I have put through so much but who never stopped cheering me on and being proud of me; my wonderful, feisty grandma Flora who sent me 21 kisses on my 21st birthday (some time ago now) and whom I miss very much; my great aunt Lexie, who enabled me to do my MSc in Biomedical Research Sciences; my brother Euan, who set the standards high and gave me someone to look up to, and my wonderful, bright, inspiring friends and office buddies. A special thank you to my wonderful husband Jody, who was my boyfriend at the start of this PhD. I know I am a nightmare at times but I know how lucky I am to have you there supporting me, believing in me and telling me to “just man the f***K up and get it done”.

Finally, thank you Alfie.

Publications

Beynon, R. A., Lang, S., Schimansky, S., Penfold, C. M., Waylen, A., Thomas, S. J., Pawlita, M., Tim, W., Martin, R. M., May, M. &**,** 2018. Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer**.** *International Journal of Cancer*. 143, 5, p.1114-1127**.**

Schimansky, S., Lang, S., **Beynon, R**., Penfold, C., Davies, A., Waylen, A., Thomas, S., Pring, M., Pawlita, M., Waterboer, T. & Ness, A., 2019. Association between comorbidity and survival in head and neck cancer: Results from Head and Neck 5000**.** *Head and Neck*. 41, 4, p.1053-1062**.**

Lang, S., Schimansky, S., **Beynon, R**., Penfold, C., Davies, A., Waylen, A., Thomas, S., Pring, M., Pawlita, M., Waterboer, T. & Ness, A. R., 2019**.** Dietary behaviors and survival in people with head and neck cancer: Results from Head and Neck 5000**.** *Head and Neck*. 41, 7, p.2074-2084**.**

Davies, A., Waylen, A., Leary, S., Thomas, S., Pring, M., Janssen, B., **Beynon, R.**, Lang, S., Schimansky, S., Hurley, K. & Ness, A., 2020**.** Assessing the validity of EQ-5D-5L in people with head & neck cancer: Does a generic quality of life measure perform as well as a disease-specific measure in a patient population? *Oral Oncology*. 101, 104504.

Langdon, RJ*., **Beynon, RA***., Ingarfield, K., Marioni, RE., McCartney, DM., Richard, MM., Ness, AR., Pawlita, M., Waterboer, T., Relton, C., Thomas, SJ., Richmond, RC., 2020. Epigenetic prediction of complex traits and mortality in a cohort of individuals with oropharyngeal cancer*. Clinical Epigenetics*. 12:58.
*These authors contributed to the manuscript equally.

Pre-prints:
Beynon, RA., Ingle, S., Langdon, R., May, M., Ness, A., Martin, RM., Suderman, M., Ingarfield, K., Marioni, R., McCartney, D., Waterboer, T., Pawlita, M., Relton, C., Davey Smith, G., Richmond, RC., 2020. Epigenetic biomarkers of ageing are predictive of mortality risk in a longitudinal clinical cohort of individuals diagnosed with oropharyngeal cancer. *MedRxiv doi*: https://doi.org/10.1101/2020.02.04.20020198.

Table of contents

List of Appendices **21**

List of abbreviations **24**

Chapter 1: Thesis overview **28**

1.1. Thesis motivation and aim *28*

1.2. Organisation *28*

Chapter 2: Introduction to head and neck cancer **30**

2.1. What is head and neck cancer? *30*

2.2. Head and neck cancer sub-sites *30*

2.2.1. Oral cavity 30

2.2.2. Nasopharynx 30

2.2.3. Oropharynx 31

2.2.4. Larynx 31

2.2.5. Hypopharynx 31

2.2.6. Nose and paranasal sinuses 31

2.3. ICD codes *34*

2.4. How are head and neck cancers diagnosed? *38*

2.4.1. Signs and symptoms 38

2.4.2. History and physical exam 38

2.4.3. Diagnostic tests 38

2.5. Staging of head and neck cancer *39*

2.5.1. TNM staging 40

2.5.2. Prognostic staging 45

2.5.3. Pros and cons of the TNM system 46

2.6. Head and neck cancer treatment *47*

2.6.1. Treating early stage cancers 47

2.6.2. Treating locally advanced cancer 48

2.7. Summary *49*

Chapter 3: Head and neck cancer epidemiology, biomarkers and prognosis **50**

3.1. Introduction *50*

3.2. Prevalence of head and neck cancer 50

3.3. HNC Incidence rates 51

3.3.1. UK incidence rates 51

3.3.2. Global variations in incidence rates 54

3.3.2.1. Lip and oral cavity cancer 55

3.3.2.2. Oropharyngeal cancer 59

3.3.2.3. Laryngeal cancer 59

3.3.2.4. Other HNC sites 60

3.4. Trends in head and neck cancer incidence rates 60

3.4.1. UK trends in incidence rates 61

3.4.2. Global trends in incidence rates 62

3.5. Mortality rates 64

3.5.1. UK mortality rates 64

3.6. Major risk factors for HNC 66

3.6.1. Tobacco use 67

3.6.2. Alcohol use 69

3.6.3. Human papilloma virus infection 70

3.6.4. Epstein-Barr virus 73

3.6.5. Diet and nutrition 73

3.6.6. Body mass index 73

3.6.7. Physical activity 74

3.6.8. Oral hygiene 74

3.6.9. Socioeconomic position 75

3.6.10. Family history of head and neck cancer 75

3.6.11. Genetic risk factors 76

3.7. Multi-omics profiling in HNC 81

3.7.1. Genetic landscape 81

3.7.2. Epigenetic signatures 82

3.7.3. Metabolomic profiles 87

3.8. Survival rates after HNC diagnosis 93

3.9. Head and neck cancer prognostic factors 94

3.9.1. Individual-related HNC prognostic factors 94

3.9.1.1. Age 94

3.9.1.2. Gender 95

3.9.1.3. Race/ethnicity 95
3.9.1.4. Socioeconomic position 95
3.9.1.5. Comorbidity 96
3.9.2. Tumour-related HNC prognostic factors 97
3.9.2.1. Tumour stage 97
3.9.2.2. HPV status 98
3.9.2.3. Genetic predictors 98
3.9.2.4. Epigenetic predictors 98
3.9.2.5. Metabolic predictors 99
3.9.2.6. Infection biomarkers 99
3.9.2.7. Other molecular biomarkers 100
3.9.3. Environment-related HNC prognostic factors 100
3.9.3.1. Smoking 100
3.9.3.2. Alcohol drinking 101
3.9.3.3. Diet 101
3.9.4. HNC prognostic models 107
3.10. Summary 115

Chapter 4: Capturing exposures to biological, environmental and lifestyle risk factors 116

4.1. Introduction 116
4.2. Self-reported phenotypes 116
4.3. Biochemical measures of tobacco and alcohol exposure 118
4.4. Epigenetic predictors of health and lifestyle 118
4.4.1. What is DNA methylation? 118
4.4.2. DNAm-based predictors of smoking, alcohol consumption, educational attainment and BMI 122
4.4.3. Epigenetic biomarkers of aging 123
4.4.3.1. Single-tissue DNA methylation-based age estimators 124
4.4.3.2. Multi-tissue DNA-methylation based age estimators 125
4.4.3.3. Phenotypic age estimator 125
4.4.3.4. GrimAge 126
4.4.3.5. Epigenetic age acceleration 127
4.5. HPV detection methods in HNC 128
4.5.1. p16 immunohistochemistry 129
4.5.2. HPV in situ hybridization 130

4.5.3. HVP polymerase chain reaction testing 130
4.5.4. HPV serological testing 131
4.6. Metabolomic assessment of exposure 137
4.7. Summary of the challenges and opportunities 140
Chapter 5: The Head and Neck 5000 study 142
5.1. Study design and follow-up 142
5.1.1. Inclusion criteria: 142
5.1.2. Exclusion criteria: 142
5.1.3. Consent 143
5.1.4. Recruitment rates 143
5.1.5. Baseline data collection 143
5.1.6. Response rates 145
5.1.7. Follow-up 145
5.2. Methods for collecting and processing materials 147
5.2.1. Blood sample collection, processing, and storage 147
5.2.2. Genotyping and imputation 147
5.2.3. DNA methylation profile generation 148
5.2.4. Metabolite quantification 149
5.2.5. Determination of HPV status 151
5.3. Variable description 152
5.3.1. Demographic variables 152
5.3.1.1. Gender 152
5.3.1.2. Age at consent 153
5.3.1.3. Ethnicity 153
5.3.2. Clinical variables 154
5.3.2.1. Primary Diagnosis 154
5.3.2.2. TNM staging 156
5.3.2.3. HPV status 157
5.3.2.4. Body mass index 158
5.3.2.5. Comorbidity 159
5.3.2.6. Cancer care plan intent 160
5.3.3. Socioeconomic variables 160
5.3.3.1. Annual household income 160
5.3.3.2. Highest education level obtained 161
5.3.3.3. Marital status 162

5.3.4. Lifestyle behaviour variables 163
5.3.4.1. Alcohol consumption 163
5.3.4.2. Baseline smoking status 164
5.4. Representativeness of the cohort 165
Chapter 6: A description of the datasets used in the thesis 167
6.1. Introduction 167
6.2. Participants included in the observational analysis 168
6.3. Participants included in the epigenetic analyses 171
6.4. Participants included in the metabolomics analysis 174
6.5. Discussion 177
Chapter 7: Associations of self-reported smoking status and alcohol use at diagnosis with survival in H&N5000 178
7.1. Introduction 178
7.2. Aims and objectives 179
7.3. Methods 179
7.3.1. Study population 179
7.3.2. Outcome assessment 180
7.3.3. Defining exposures 180
7.3.4. Assessment of HPV status 180
7.3.5. Statistical analysis 181
7.3.5.1. Descriptive analysis 181
7.3.5.2. Accounting for missing data 181
7.3.5.3. Survival analysis 182
7.4. Results 184
7.4.1. Missing data 184
7.4.2. Baseline characteristics of study population 187
7.4.3. Kaplan-Meier survival plots 189
7.4.4. Variation in survival explained by smoking status and alcohol intake 192
7.4.5. Smoking status and survival 193
7.4.6. Alcohol intake and survival 193
7.4.7. Heterogeneity between centres 194
7.4.8. Influence of tumour stage on the associations of smoking and alcohol intake with survival 194

7.4.9. Influence of HPV status on the associations of smoking and alcohol intake with survival 195
7.4.10. Interaction of tobacco and alcohol 196
7.5. Discussion 200
7.5.1. Principal findings 200
7.5.2. Strengths and limitations of the study 201
7.5.3. Conclusions 203

Chapter 8: Epigenetic prediction of complex traits and mortality in oropharyngeal cancer in H&N5000 204
8.1. Introduction 204
8.2. Aims and objectives 205
8.3. Methods 205
8.3.1. Study population 205
8.3.2. Assessment of tobacco, alcohol, BMI, and education 205
8.3.3. Epigenetic profiling and pre-processing 206
8.3.4. Epigenetic risk score generation 206
8.3.5. Multiple imputation 207
Statistical analysis 210
8.3.5.1. Associations of epigenetic scores with self-reported phenotypes 210
8.3.5.2. Proportion of variance in survival explained by DNAm scores 210
8.3.5.3. Survival analysis 210
8.4. Results 212
8.4.1. Baseline characteristics of the study population 212
8.4.2. Correlation between covariates 216
8.4.3. Correlation between DNAm predictors 216
8.4.4. Proportion of variance in phenotype explained by the DNAm scores 216
8.4.6. Proportion of variance in survival explained by the DNAm scores 219
8.4.7. Association of DNAm predictors with all-cause mortality 220
8.5. Discussion 225
8.5.1. Principle findings 225
8.5.2. Strengths and limitations of the study 226
8.5.3. Conclusions 227

Chapter 9: Associations of epigenetic biomarkers of ageing with mortality risk in oropharyngeal cancer 228

9.1. Introduction 228

9.2. Aims and objectives 229

9.3. Methods 230

9.3.1. Study population 230

9.3.2. Estimation of epigenetic age 230

9.3.3. Statistical analysis 233

9.3.3.1. Step 1: Examining the association of EAA measures with survival 233

9.3.3.2. Step 2: Assessing the prognostic value of EAA measures 234

9.4. Results 237

9.4.1. Baseline descriptives 237

9.4.2. Pairwise correlations between measures of epigenetic age acceleration 238

9.4.3. Explained variation in survival 240

9.4.4. Association of DNA Methylation-Based Biological Age with survival 240

9.5. Discussion 247

9.5.1. Principle findings 247

9.5.2. Strengths and limitations of the study 249

9.5.3. Conclusions 249

Chapter 10: Metabolic signatures of oropharyngeal cancer survival 251

10.1. Introduction 251

10.2. Aims and objectives 252

10.3. Methods 252

10.3.1. Study population 253

10.3.2. Measurement of metabolites 253

10.3.3. Issues of multiple testing 258

10.3.3. Missing data decisions 261

10.3.4. Statistical analyses 263

10.3.5. Sensitivity analysis 264

10.4. Results 265

10.4.1. Baseline characteristics 265

10.4.2. Differences in metabolic trait concentrations between HPV-positive and HPV-negative individuals 265

10.4.3. Associations of pre-treatment metabolic traits with all-cause mortality 266

10.4.4. Sensitivity analyses 271

10.5. Discussion 273
10.5.1. Principle findings 273
10.5.2. Challenges 275
10.5.3. Strengths and limitations of the study 276
10.5.4. Conclusion 278

Chapter 11: Discussion 279
11.1. Introduction 279
11.2. Summary of findings and implications 279
11.3. Strengths and limitations of this thesis 284
11.3.1. Strengths 284
11.3.1.1. Prospective study design 284
11.3.1.2. Availability of baseline clinical, biological and lifestyle data. 284
11.3.1.3. Data linkage 285
11.3.1.4. Sample size relative to other HNC studies 285
11.3.2. Weaknesses 286
11.3.2.1. Selection into the H&N5000 study 286
11.3.2.2. Selection into the analytic sample 286
11.3.2.3. Missing covariate data 287
11.3.2.4. Power and sample size 288
11.3.2.5. Lack of cancer-specific mortality data 289
11.3.2.6. Sample collection and handling 289
11.3.2.7. Limited epigenetic and metabolomic data 290
11.3.2.8. External validity 290
11.4. Future work 291
11.4.1. Mendelian randomization study of metabolite profiles 291
11.4.2. Saliva and tissue metabolites and survival in HNC 292
11.4.3. HNC prognostic model development 293
11.4.4 Application and impact of a new HNC prognostic model in clinical practice 293
11.5. Concluding remarks 295

References: 386

List of Tables

Table 1: ICD-10 diagnosis codes for cancers of the head and neck. **35**

Table 2a: Tumour (T) staging according to anatomical site. **41**

Table 2b: Classification of lymph nodes by anatomical site **44**

Table 3: AJCC (8th edition) prognostic stage groups for HPV-associated (p16+) OPC. **47**

Table 4: Estimated number of prevalent cases of HNC in the UK (2018) as a proportion, ages 20-85+ years. **51**

Table 5: UK estimated age-standardised incidence rates of head and neck cancer by site and gender, all ages (2018). **52**

Table 6: Global estimated age-standardized incidence rates of head and neck cancer by site and gender, all ages (2018). **55**

Table 7: Some of the most intensely studies genetic polymorphisms linked to HNC risk. **78**

Table 8: HNC risk loci identified in GWAS. **80**

Table 9: Frequency of selected genes recurrently mutated in HNC. **85**

Table 10: Genes frequently hypermethylated in HNC. **86**

Table 11: Summary of metabolomic-based studies on HNC. **90**

Table 12: Head and neck survival by sub-site for males and females. **94**

Table 13: The association of pre-treatment smoking with head and neck cancer outcomes. **102**

Table 14: Association of pre-treatment alcohol consumption with head and neck cancer outcomes. **105**

Table 15: Summary description of HNC prognostic models described in the literature. **112**

Table 16: A summary of some of the characteristics of available alcohol biomarkers, Adapted from [490] and [499]. **120**

Table 17: HPV detection methods in HNC. **134**

Table 18: A comparison of the advantages and disadvantages of the two most common techniques used in metabolomics data acquisition. **139**

Table 19: Baseline characteristics of the study sample, stratified by tumour site (n=1,403). **169**

Table 20: Baseline characteristics of the individuals with oropharyngeal tumours, stratified by HPV status (n=656). **170**

Table 21: Baseline characteristics of all participants included in the epigenetic analyses (n=408). **172**

Table 22: baseline characteristics of participants included in the epigenetic analysis, stratified by HPV status (n=408). **173**

Table 23: baseline characteristics of the study sample included in the metabolomic analysis (n=703). **175**

Table 24: baseline characteristics of the study sample, stratified by HPV status (n=703). **176**

Table 25: Proportion of missing data, overall and stratified by tumour site. **186**

Table 26: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by tumour site. **197**

Table 27: Association of smoking status and alcohol intake with mortality risk, stratified by tumour stage. **198**

Table 28: Association of smoking status and alcohol intake with mortality risk, stratified by HPV status. **199**

Table 29: An overview of the DNAm scores employed in the current analysis. **208**

Table 30: Proportion of missing data in the epigenetic dataset (n=408). **210**

Table 31: Average DNAm values for people who are alive at 3-years compared to people who are dead at 3-years. **214**

Table 32: Pearson's correlation coefficient matrix for included covariates. **217**

Table 33: Pearson's correlation coefficient matrix for DNAm scores. **218**

Table 34: The proportion of phenotypic variance explained in each trait by the respective DNAm-based predictor. **219**

Table 35: Overview of various measures of epigenetic age acceleration. **232**

Table 36: Baseline characteristics of the study sample stratified by 3-year mortality status (n=408). **239**

Table 37: Measures of model performance. **245**

Table 38: Estimated coefficients (uncorrected and corrected) for the clinical + AgeAccelGrim model **247**

Table 39: An overview of the metabolic biomarkers included in the current analysis (n=145). **.....255**

Table 40: The impact on effect estimates of removing potential outliers from the dataset, using different outlier-detection methods. **.....272**

Table 41. An overview of the research questions and findings of this thesis. **.....280**

Table 42: Regression coefficients for the predictors selected by MFP. **.....284**

List of Figures

Figure 1: Research questions addressed in this thesis. 29

Figure 2: Anatomy of the Head and Neck. 32

Figure 3: Anatomic regions of the oral cavity. 32

Figure 4: Anatomy of the oropharynx. 33

Figure 5: Areas where laryngeal cancer may form or spread. 33

Figure 6: ICD code structure. 35

Figure 7: Geographical distribution of oral cavity, oropharyngeal and laryngeal cancer incidences, UK (2007-2009). 53

Figure 8: Average number of incident cases of HNC per year and age-standardised incidence rates per 100,000 persons in the UK (2013-2015). 54

Figure 9: Global incidence of HNC, by sub-site. 56

Figure 10: Total alcohol per capita consumption (15+ years; in litres of pure alcohol), 2016. 58

Figure 11: Global prevalence of tobacco smoking among persons aged 15 years and older (2015). 58

Figure 12: Rise in HNC incidence in the UK, 1993-2017. 62

Figure 13: Trend in UK smoking rates among males and females, 1950-2010. 62

Figure 14: Head and neck Cancer mortality trends over time in the UK (1971-2016). 65

Figure 15: Head and neck cancer European age-standardised mortality rates, by age, in UK males (1971-2016). 65

Figure 16: Head and neck cancer European age-standardised mortality rates, by age, in UK females (1971-2016). 66

Figure 17: HPV-16 genome 72

Figure 18: Image of web-based calculator developed by Emerick et al. 111

Figure 19: Regulation of transcription by DNAm 119

Figure 20: Microscope images of HPV-positive and HPV-negative oropharyngeal tumours by ISH and p16 IH. 131

Figure 21: A comparison of antibody-capture enzyme-linked immunosorbent assay (left) and bead-based multiplex serology (right). 133

Figure 22: The 76 H&N5000 Study centres [574]. 146

Figure 23: H&N5000 data collection points. 146

Figure 24: Proportion of male and female participants in H&N5000. 152

Figure 25: Histogram showing age distribution in H&N5000. 153

Figure 26: Bar chart of ethnicity in HN5000. 154

Figure 27: Proportion of HNCs by sub-site in H&N5000. 156

Figure 28: Bar chart showing the proportion of HNCs in each TNM stage group in H&N5000. 157

Figure 29: Bar chart showing the proportion of HN5000 participants who were HPV-positive or HPV negative, as determined by HPV-16 seropositivity. 158

Figure 30: Histogram of BMI distribution in H&N5000. 159

Figure 31: Bar chart of showing the comorbidity burden of participants in H&N5000. 159

Figure 32: Bar chart showing the intended treatment plan for participants at baseline in H&N5000. 160

Figure 33: Bar chart of participant income in H&N5000. 161

Figure 34: Bar chart showing the education structure of respondents in H&N5000. 162

Figure 35: Bar chart showing H&N5000 participants' marital status. 163

Figure 36: bar chart showing the alcohol consumption levels of participants in H&N5000. 164

Figure 37: Bar chart showing the proportion of never-, former- and current smokers in H&N5000. 165

Figure 38: number of participants included in each of the primary analysis conducted in this thesis. 167

Figure 39: Flow of Head and Neck 5000 participants through the study. 185

Figure 40: Kaplan-Meier plot of overall survival by HNC site. 186

Figure 41: Kaplan-Meier plot of overall survival by smoking status and alcohol consumption, stratified by tumour site. 191

Figure 42: Kaplan-Meier plot of overall survival by smoking status and alcohol consumption, stratified by HPV status. 192

Figure 43: Flow of participants included in the analysis. 207

Figure 44: Forest plots showing the estimated hazard ratios and corresponding 95% confidence intervals for the associations of DNAm predictors with all-cause mortality (n=408).....223

Figure 45: Flow of participants included in the Epigenetic Age analysis.....236

Figure 46: Heat map showing the pairwise correlation coefficients between epigenetic measures of age acceleration. 238

Figure 47: Association of epigenetic age acceleration measures with mortality risk. 242

Figure 48: Independent contribution of AgeAccelGrim to prognosis beyond clinical factors (n=408).....246

Figure 49: A comparison of the area under the ROC curves (AUC) obtained for the models included in the sensitivity analyses (n=384).....246

Figure 50: Scree plot and Stata output showing the decreasing rate at which variance is explained by additional principal components.....260

Figure 51: Loading to PC1 and PC2. 261

Figure 52: OPC samples available for analysis 262

Figure 53: Cox regression results for models 1 and 2 267

Figure 54: Cox regression results for models 3 and 4. 269

Figure 55: Linear fit between complete case and full dataset models.....270

List of Appendices

List of Tables **15**

Appendix A **296**

A1: A comparison of the baseline characteristics of people with and without BMI data available (n=3,890). *286*

A2: Baseline descriptives of people included in the imputed analysis (n=3,890). *296*

A3: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by tumour site (n=3,890). *297*

A4: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by tumour stage. *298*

A5: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by HPV status. *299*

Appendix B **300**

B1: Kaplan-Meier plots of overall survival in univariate models (n=408). *300*

B2: Histograms showing the distribution of AHRR DNAm scores, overall and by self-reported smoking status (n=408). *303*

B3: Histograms showing the distribution of Joehanes (Bonferroni) DNAm scores, overall and by self-reported smoking status. *303*

B4: Histograms showing the distribution of Joehanes (FDR) DNAm scores, overall and by self-reported smoking status. *304*

B5: Histograms showing the distribution of Zhang DNAm scores, overall and by self-reported smoking status. *304*

B6: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported smoking status. *305*

B7: Histograms showing the distribution of Lui DNAm scores (based on 5 CpGs), overall and by self-reported alcohol consumption. *305*

B8: Histograms showing the distribution of Lui (23 CpG) DNAm scores, overall and by self-reported alcohol consumption. *306*

B9: Histograms showing the distribution of Lui (78 CpG) DNAm scores, overall and by self-reported alcohol consumption. *306*

B10: Histograms showing the distribution of Lui (144 CpG) DNAm scores, overall and by self-reported alcohol consumption. 307

B11: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported alcohol consumption. 307

B12: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported educational attainment level. 308

B13: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported BMI. 308

B14: Stata code used to standardise DNAm scores, to allow direct comparison across scores with different scales. 309

B15: Kaplan-Meier observed survival curves and cox predicted survival for the variable HPV status. 310

B16: Log-log plot testing the PH assumption for the variable HPV status. 310

B17: Kaplan-Meier observed survival curves and cox predicted survival for the variable comorbidity. 311

B18: Log-log plot testing the PH assumption for the variable Comorbidity. 311

B19: A comparison of the effect estimates obtained for the associations of DNAm scores with survival in the primary and complete case analyses. 312

B20: Results of the sensitivity analysis, censoring data at 3-years, adjusted for age, gender, cell counts and batch effects. 314

B21: Results of the sensitivity analysis, additionally adjusted for TNM stage, HPV status and comorbidity. 314

B22: Results of the sensitivity analysis, additionally adjusted for the corresponding directly measured phenotype. 315

B23: Results of the sensitivity analysis, additionally adjusted for the other directly measured phenotypes. 315

Appendix C 316

C1: Baseline characteristics of participants included in the complete case analysis. 316

C2: Results of the complete case cox regression analysis (n=225). 317

Appendix D 319

D1: Abbreviations, names and units of metabolic measures quantified on the Nightingale NMR platform. *319*

D2: PC loadings for the metabolic trait measures. *326*

D3: Proportion of missing data in the metabolomics dataset (n=1,483). *330*

D4: Density histograms showing the observed data distributions of metabolic trait measures (n=703). *331*

D5: Standardised mean difference in circulating metabolic trait concentrations for HPV (+) vs. HPV (-) OPCs. *380*

D6: Summary of Cox PH regression results, showing only those metabolites that reached the threshold for multiple testing (n=703). *384*

List of abbreviations

ACE-27	Adult comorbidity evaluation
ADH	Alcohol dehydrogenase
AF	Attributable fractions
AIC	Akaike Information Criterion
AJCC	American Joint Committee on Cancer
ALAT	Alanine aminotransferase
ALDH	Aldehyde dehydrogenase
ANOVA	Analysis of variance
ASAT	Aspartate aminotransferase
ASR	Age-standardised rate
ASIR	Age-standardised instance rate
AUC	Area under the receiver operating curve
AUDIT	Alcohol Use Disorders Identification Test
BIC	Bayesian Information Criterion
BMI	Body mass index
CCI	Charlson Comorbidity Index
CDK	Cyclin-dependent kinase
CDT	Carbohydrate-deficient transferrin
TCGA	The Cancer Genome Atlas
CHARGE	Heart and aging Research in Genetic Epidemiology Consortium
CI	Confidence interval
CpG	Cytosine-guanine dinucleotide
CRUK	Cancer Research UK
CT	Computerised tomography
cTNM	Clinical TNM staging (see TNM staging)
DCF	Data capture form
DFS	Disease-free survival
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
DNMT	DNA methyltransferase
dNTP	Deoxyribonucleotide triphosphate
ddNTPs	Dideoxynucleotides triphosphates
DSS	Disease-specific survival
EAA	Epigenetic age acceleration
EEAA	Extrinsic epigenetic age acceleration
EBRT	External beam radiotherapy
EBNA1	Epstein-Barr nuclear antigen 1
EBV	Epstein-Barr virus
ECS	Extra capsular spread
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
ELISA	Enzyme Linked Immunosorbent Assay
EOR	Excess odds ratio

EPIC	European Prospective Investigation into Cancer and Nutrition
ER	Estrogen receptor
ER	Endoplasmic reticulum
ESP	European standard population
EtG	Ethyl glucuronide
EU	European Union
EUCAN	European Cancer Observatory
E6AP	E6-associated protein
FAEE	Fatty acid ethyl esters
FDG	Fluorodeoxyglucose
FDR	False Discovery Rate
FFPE	Formalin fixed paraffin embedded
FNAC	Fine needle aspiration cytology
GGT	Gamma–glutamyltransferase
GST	Glutathione S-transferase
Gy	Grays
GWAS	Genome wide association study
HDL	High density lipoprotein
HDI	Human development index
HER2	Human epidermal growth factor receptor 2
Hex	Hexosaminidase
His	Histidine
HNC	Head and neck cancer
H&N5000	Head and Neck 5000
HNSCC	Head ans Neck squamous cell carcinoma
H&P	History and physical exam
HPV	Human papilloma-virus
HR	Hazard ratio
HRF	Haplotype reference panel
HSCIC	Health and Social care information Centre
IARC	International Agency for Research into Cancer
ICD	The International Classification of Diseases and Related Health Problems
IDL	Intermediate density lipoprotein
IEAA	Intrinsic epigenetic age acceleration
IEU	Integrative Epidemiology Unit
IHC	Immunohistochemistry
IMD	Index of Multiple Deprivation
IMRT	Intensity-modulated radiation therapy
INHANCE	International Head and Neck Cancer Epidemiology consortium
IQR	Interquartile range
ISH	In-situ hybridisation
LASSO	Least absolute shrinkage
LCR	Long control region
LDL	Low density lipoprotein
IL6/7	Interleukin 6/7

LMP-1	Latent membrane protein 1
LMR	Lymphocyte-to-monocyte ratio
LMWM	Low molecular weight molecules
LNM	Lymph node metastasis
LSOA	Lower Layer Super Output Area
mAB	Monoclonal antibody
MAR	Missing at random
MCAR	Missing completely at random
MDT	Multi-disciplinary team
MFI	Median fluorescence intensity
MI	Multiple imputation
MNAR	Missing not at random
MPG	Multiplex HPV genotyping
MRI	Magnetic resonance imaging
MS	Mass spectrometry
NADH	Nicotinamide adenine dinucleotide
NCIN	National Cancer Intelligence Network
NF-kB	Nuclear factor kappa B
NHANES	National Health and Nutrition Examination Survey
NHS	National Health Service
NHSCR	NHS Central Register
NHSIC	National Health Service Information Centre
ncRNA	Noncoding ribonucleic acids
NICE	National institute for clinical excellence
NK	Natural killer cells
NLR	Neutrophil-to-monocyte ratio
NMR	Nuclear magnetic resonance
NPC	Nasopharyngeal cancer/carcinoma
NPI	Nottingham prognostic index
OCC	Oral cavity cancer
OCIN	Oxford Cancer Intelligence Network
OLK	Oral leukoplakia
ONS	Office for national statistics
OPC	Oropharyngeal cancer
ORF	Open reading frame
ORF	Odds ratio
OS	Overall survival
OSSC	Oral squamous cell carcinoma
Ossig	Overall survival signature
PA	Physical activity
PAH	Polynuclear aromatic hydrocarbon
PAR	Population attributable risk
PCR	Polymerase chain reaction
PET	Positron emission tomography
PH	proportional hazards
PLR	platelet-to-lymphocyte ratio

pTNM	pathological TNM staging [see TNM staging]
QC	Quality control
QoL	Quality of life
ROS	Reactive oxygen species
RR	Rate ratio
RT–PCR	Reverse transcriptase PCR
SA	Sialic acid
SCC	Squamous cell carcinoma
SEER	Surveillance, Epidemiology, and End Results
SES	Socioeconomic status
SIR	Standardised incidence ratio
SMC	Squared multiple correlations
SNP	single nucleotide polymorphism
SPT	Second primary tumour
TF	Transcription factor
TNF-α	Transforming growth factor alpha
TNM	Tumour node metastasis staging
TS	Tumour suppressor
T-stage	Describes the size of the primary tumour
UADT	Upper aerodigestive tract
UICC	Union for International Cancer Control
UK	United Kingdom
US	United states
VEGF	Vascular endothelial growth factor
WGA	Whole genome amplification
WHO	World Health Organisation

Chapter 1: Thesis overview

1.1. Thesis motivation and aim

Assessing prognosis is a major component of the clinical encounter. For someone with a diagnosis of head and neck cancer (HNC), the outcome of their disease is a major concern. The challenge for the clinician is to predict the course of the cancer (e.g. the likelihood of metastasis or death in a given timeframe) in a particular individual, in order to inform treatment decisions.

Prognostic models (also known as clinical prediction rules) can help clinicians to both predict an individual's likely clinical outcome and discuss their prognosis with them. Existing HNC prognostic models which, in routine practice, are based almost exclusively on clinical biomarkers, such as TNM stage and tumour site, predict mortality with varying success. For this reason, coupled with the growing ability to capture and mine vast amounts of clinical and biological data on individuals, the study of prognostic factors is increasingly important.

The role of certain lifestyle factors such as smoking and alcohol intake in HNC prognosis have been investigated. However, measures of such exposures are prone to error because they often rely on self-report, leading to potentially inaccurate estimates of the effect of these exposures on HNC outcomes. In addition, many studies looking at potential prognostic factors have been small, and therefore underpowered, or they have lacked data on important confounders.

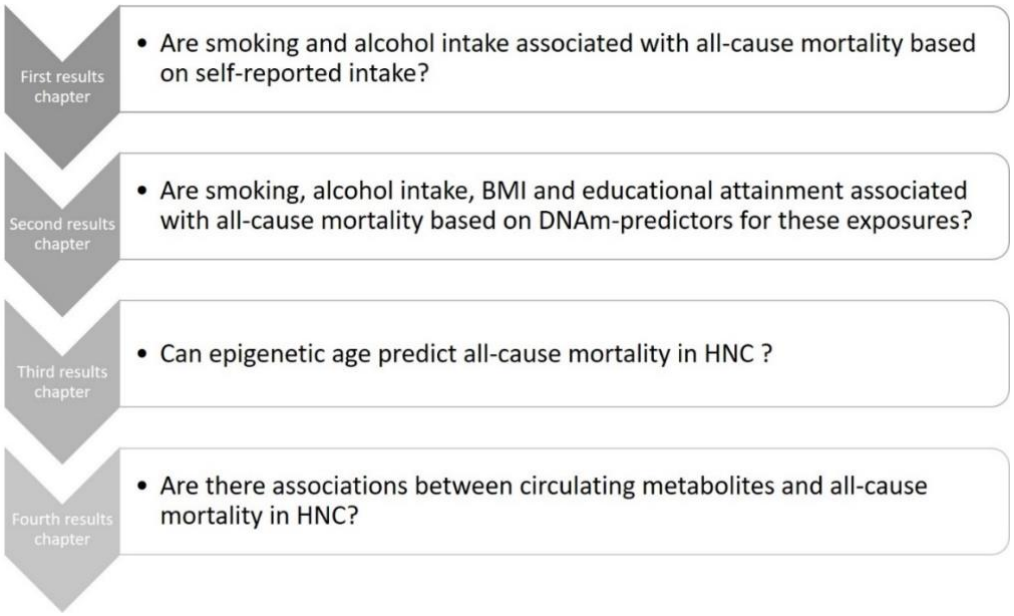
The overall aim of this thesis is to investigate, using a combination of self-reported phenotypes, DNA methylation-based predictors of exposures, and information on circulating metabolites, whether specific lifestyle traits and life-course exposures predict all-cause mortality and should be considered for inclusion in HNC prognostic models. The findings of this analysis could enhance clinical decision making in this population and help stratify people with disease in future epidemiological studies and clinical trials.

1.2. Organisation

This thesis starts by describing the biological and clinical aspects of HNC and provides an overview of its epidemiology. It then summarises the existing prognostic literature in this

area and some of the major challenges encountered when trying to measure lifestyle exposures that could potentially serve as prognostic biomarkers. In the next section, I describe the clinical cohort study from which the data for this PhD was obtained. I provide a brief overview of the data collection methods and the baseline characteristics of the overall cohort, before describing the specific analytic datasets used in my analysis. The main research questions addressed in the subsequent results chapters are presented in Figure 1.

Figure 1: Research questions addressed in this thesis.



The first results chapter examines the potential associations of self-reported smoking and alcohol intake with all-cause mortality, using data collected on a subset of Head and Neck 5000 (H&N5000) participants with cancers of the oral cavity, oropharynx and larynx. Given the known limitations of self-reported lifestyle data, the second results chapter investigates the strength of the associations between these exposures and other complex traits including body mass index (BMI) and educational attainment, with all-cause mortality, using DNA-methylation (DNAm) based predictors for these traits. Here, the analysis is restricted to cohort members with oropharyngeal cancer (OPC) only. The next two results chapters also focus on OPC, specifically. The third results chapter examines the predictive utility of DNAm-based predictors of biological aging on all-cause mortality. The fourth results chapter explores the possible association between participants' circulating metabolic profiles and all-cause mortality. A summary of my main findings and potential future work is provided in the final discussion chapter.

Chapter 2: Introduction to head and neck cancer

2.1. What is head and neck cancer?

HNC is an umbrella term used to describe malignancies at several different sites in the head and neck region. These malignancies include cancers of the mouth, throat (larynx), thyroid, nasal cavity, and sinuses. Skin cancers and brain tumours are not usually included under this definition [1,2]. Around 90% of HNCs are squamous cell carcinomas (SCC) [3], meaning that the arise in the thin, flat cells that line the surfaces of these structures. Other types of HNCs include adenocarcinomas (cancers that originate in the mucous glands), melanomas (cancers that develop from melanocytes in the epidermis), lymphomas (cancers that start in the lymph nodes or lymph gland tissue), and sarcomas (cancers that develop in the connective tissue e.g., in the bone, fat, muscle or cartilage) [4]. The term "head and neck cancer" as discussed here will be restricted to SCC as they make up the majority of cases and because they share common risk factors.

2.2. Head and neck cancer sub-sites

The head and neck region is one of the most complicated anatomical structures in the human body. This is because it provides the mechanisms for many basic functions, including speech, swallowing, hearing and breathing. HNCs are defined by the area in which they arise. The major sites include: the oral cavity (mouth), the nasopharynx, the oropharynx, the larynx (voice box), the hypopharynx, and the nose and paranasal sinuses (Figure 2). Each major anatomical site is further divided into several subsites, as outlined below.

2.2.1. Oral cavity

includes the mucosal surfaces of the lip, the lining inside the cheeks and floor of the mouth (buccal mucosa), the front two-thirds of the tongue, the bony structure at the top of the mouth (hard palate), and the small area of gum behind the wisdom teeth (retromolar triangle) [5] (Figure 3).

2.2.2. Nasopharynx

the upper section of the pharynx (throat), which connects the back of the nose to the back of the mouth (oropharynx) [6] (Figure 2).

2.2.3. Oropharynx

the part of the throat that includes the soft palate, base of the tongue, uvula, palatine tonsils and tonsillar pillars [6] (Figure 4).

2.2.4. Larynx

the area of the throat that is located below the oropharynx and in front of the hypopharynx [7]. It is subdivided into three subsites: the supraglottis, glottis, and subglottis (Figure 5). The supraglottic larynx includes the epiglottis (a small flap of tissue that closes off the larynx when eating to prevent food from entering the airways), the ventricular bands (false vocal cords), the arytenoids cartilages, and the aryepiglottic folds. The glottis consists of the true vocal cords and their anterior and posterior commissures. The subglottic region begins just below the true vocal cords and extends to the lower edge of the cricoid cartilage, a ring of cartilage that surrounds the trachea (windpipe) [6].

2.2.5. Hypopharynx

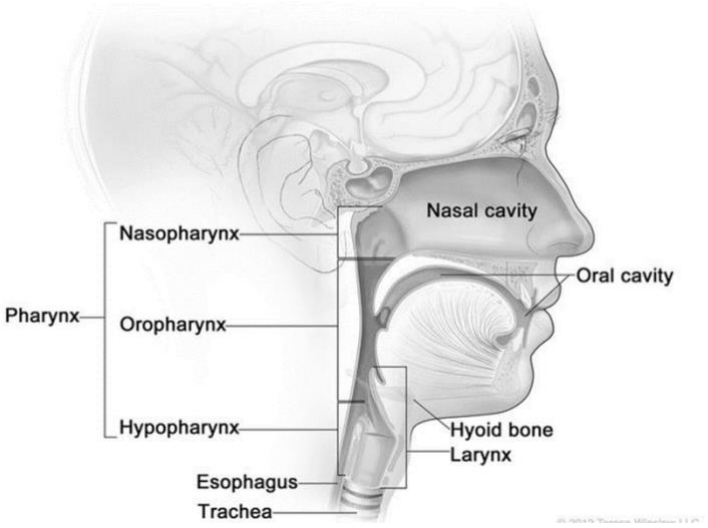
the region between the oropharynx and the oesophagus (the tube that connects the throat to the stomach; Figure 2). The major subsites of the hypopharynx include the pyriform sinuses, the post-cricoid region, and the lateral and posterior pharyngeal walls [6]. The vast majority of hypopharynx cancers arise in the pyriform sinus [8].

2.2.6. Nose and paranasal sinuses

includes the lining of the nasal cavity (medial maxillary walls), the nasal septum, and four pairs of air-filled hollows in the bones around the nose. These include the maxillary sinuses, which are located under the eyes, the superior frontal sinuses, which are located below the eyes, the bilateral ethmoid sinuses, which sit between the eyes, and the sphenoidal sinuses, located behind the eyes [9] (Figure 2).

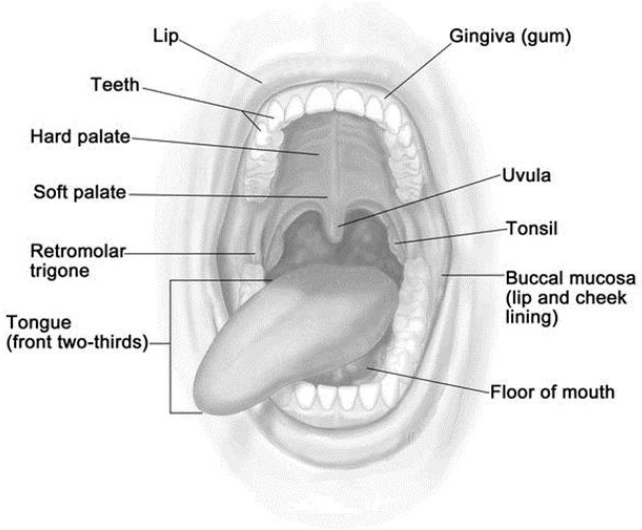
Cancers can also arise in the thyroid gland. However, SSC of the thyroid are extremely rare, representing <1% of all primary carcinomas of the thyroid [10]. Similarly, SCC only make up 0.9% to 4.7% of all major salivary gland tumours [11], therefore these cancers will not be discussed further in this thesis.

Figure 2: Anatomy of the Head and Neck.



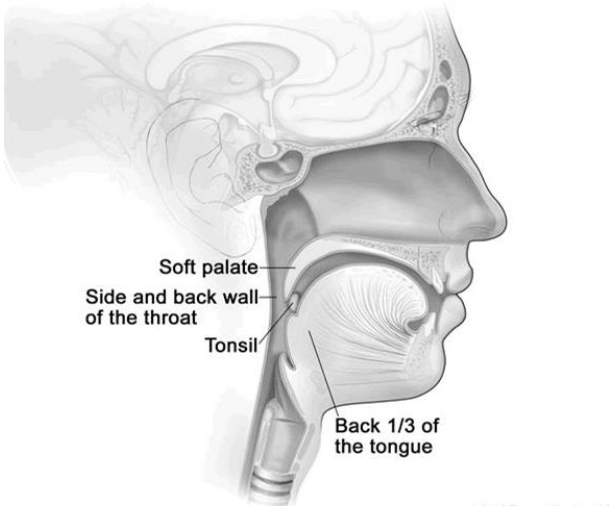
For the National Cancer Institute © 2016 Terese Winslow LLC.

Figure 3: Anatomic regions of the oral cavity.



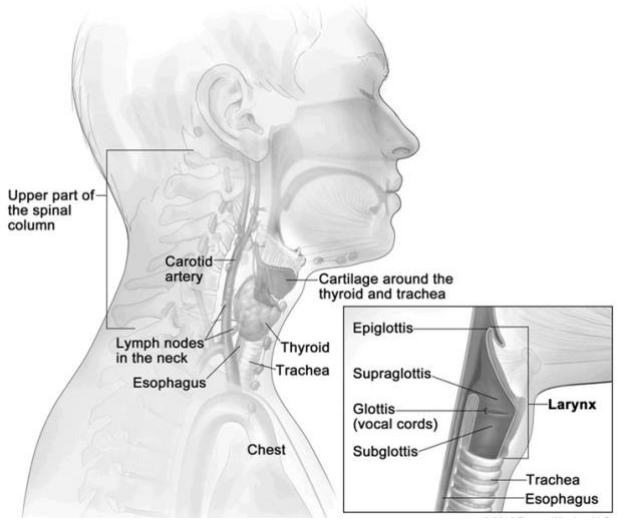
For the National Cancer Institute © 2016 Terese Winslow LLC.

Figure 4: Anatomy of the oropharynx.



For the National Cancer Institute © 2016 Terese Winslow LLC.

Figure 5: Areas where laryngeal cancer may form or spread.



For the National Cancer Institute © 2016 Terese Winslow LLC.

2.3. ICD codes

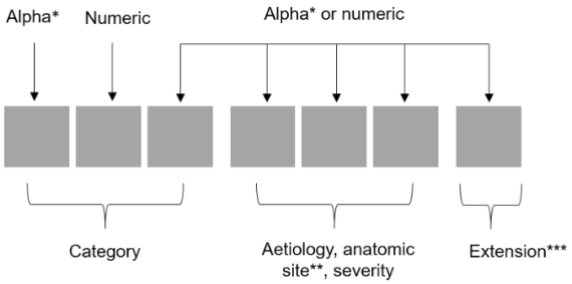
Given the complex topography and histology of HNC, standardised classification systems are necessary to register diagnoses and to facilitate the systematic analysis and comparison of morbidity and mortality data. The International Classification of Diseases and Related Health Problems, more commonly referred to by the acronym ICD, is the standard diagnostic classification tool used by physicians, policy makers and epidemiologists [12].

The ICD is currently in its tenth revision (ICD-10), although the eleventh revision (ICD-11) was released in June 2018 and will be come into effect on 1 January 2022 [13].
The ICD system is based on a unique set of alphanumeric codes. Codes in the current revision can be three, four, five, six or seven characters in length. A detailed description of the classification system along with full coding guidelines can be found in the second volume of the International statistical classification of diseases and related health problems 10th Revision (sections 2.3 and 2.4) [14]. At its simplest level, the first three-character codes, known as 'rubrics', specify the category of the diagnosis (Figure 6). The first character is always a letter. In this case, the letter "C" indicates that the diagnosis relates to a neoplasm or cancer. As an example, when "C" is used in conjunction with the numerals "0" and "3", this indicates that the diagnosis falls into the category of malignant neoplasm of the gum. Characters four to six describe in greater detail the cause, anatomic location (e.g. laterality) and severity of the disease. Using the above example, cancer of the gum (ICD-11 C03) can be subdivided into the upper gum (ICD-11 C03.0) and the lower gum (ICD-11 C03.1) using 0 or 1 in the fourth position. Not all codes include all the positions and the seventh character is typically only used for injury and external cause-related codes. Table 1 lists the ICD-10 medical diagnostic codes used for malignant neoplasms of the head and neck.

The way in which individual ICD codes (sub-sites) are grouped varies in the epidemiological literature and across Cancer registries [15]. In particular, there is considerable debate around which sites should be classified as oral cavity cancer and which sites should be classified as oropharyngeal. The decision is partly based on anatomical location but also on where HPV infection, an established risk factor for oropharyngeal cancer, is found [16 17]. Based on a review of the epidemiological literature, Conway et al propose that oropharyngeal cancer (OPC) should be defined as the sites: base of tongue (C01), lingual tonsil (C2.4), tonsil (C09), oropharynx (C10), and pharynx unspecified including Waldeyer's ring /over-lapping sites of oral cavity and pharynx (C14); while oral cavity cancer includes: the inner lip (C00.3 – C00.9), other and unspecified parts of the tongue (C02) (excluding lingual tonsil [C2.4]),

gum (C03), floor of the mouth (C04), palate (C05), and other and unspecified parts of the mouth (C06) [15]. A breakdown of the ICD-codes used to identify HNC sites in this thesis is provided in Chapter 5.

Figure 6: ICD code structure.



Every letter except "U". ** The number 1 is used to indicate right side, number 2 to indicate left side, number 3 to indicate bilateral, number 9 indicates side is unspecified in the medical record.*Required for certain codes, including 'S' codes (injuries and external causes), to provide information about the characteristic of the encounter: A= internal encounter, B=subsequent encounter, C=Sequelae (complications that arise as a direct result of a condition).*

Table 1: ICD-10 diagnosis codes for cancers of the head and neck [18].

Lip	
C00.0	External upper lip
C00.1	External lower lip
C00.2	External lip, NOS
C00.3	Mucosa of upper lip
C00.4	Mucosa of lower lip
C00.5	Mucosa of lip, NOS
C00.6	Commissure of lip
C00.8	Overlapping lesion of lip
C00.9	Lip, NOS (excludes Skin of lip C44.0)
Base of tongue	
C01.9	Base of tongue, NOS
Other unspecified parts of the tongue	
C02.0	Dorsal surface of tongue, NOS
C02.1	Border of tongue
C02.2	Ventral surface of tongue, NOS
C02.3	Anterior 2/3 of tongue, NOS
C02.4	Lingual tonsil
C02.8	Overlapping lesion of tongue
C02.9	Tongue, NOS

Table 1 continued.

Gum	
C03.0	Upper gum
C03.1	Lower gum
C03.9	Gum, NOS
Floor of the mouth	
C04.0	Anterior floor of mouth
C04.1	Lateral floor of mouth
C04.8	Overlapping lesion of floor of mouth
C04.9	Floor of mouth, NOS
Palate	
C05.0	Hard palate
C05.1	Soft palate, NOS (excludes Nasopharyngeal surface of soft palate C11.3)
C05.2	Uvula
C05.8	Overlapping lesion of palate
C05.9	Palate, NOS
Other and unspecified parts of the mouth	
C06.0	Cheek mucosa
C06.1	Vestibule of mouth
C06.2	Retromolar area
C06.8	Overlapping lesion of other and unspecified parts of mouth
C06.9	Mouth, NOS
Paratoid gland*	
C07.9	Parotid gland
Other and unspecified major salivary glands*	
C08.0	Submandibular gland
C08.1	Sublingual gland
C08.8	Overlapping lesion of major salivary glands
C08.9	Major salivary gland, NOS (excludes minor salivary gland, NOS C06.9)
Tonsil	
C09.0	Tonsillar fossa
C09.1	Tonsillar pillar
C09.8	Overlapping lesion of tonsil
C09.9	Tonsil, NOS (excludes lingual tonsil C02.4 and pharyngeal tonsil C11.1)
Oropharynx	
C10.0	Vallecula
C10.1	Anterior surface of epiglottis
C10.2	Lateral wall of oropharynx
C10.3	Posterior wall of oropharynx
C10.4	Branchial cleft (site of neoplasm)
C10.8	Overlapping lesion of oropharynx
C10.9	Oropharynx, NOS
Nasopharynx	
C11.0	Superior wall of nasopharynx
C11.1	Posterior wall of nasopharynx

Table 1 continued.

C11.2	Lateral wall of nasopharynx
C11.3	Anterior wall of nasopharynx
C11.8	Overlapping lesion of nasopharynx
C11.9	Nasopharynx, NOS
Pyriform sinus	
C12.9	Pyriform sinus
Hypopharynx	
C13.0	Postcricoid region
C13.1	Hypopharyngeal aspect of aryepiglottic fold, NOS (excludes laryngeal aspect of aryepiglottic fold C32.1)
C13.2	Posterior wall of hypopharynx
C13.8	Overlapping lesion of hypopharynx
C13.9	Hypopharynx, NOS
Other and ill-defined sites in the lip, oral cavity and pharynx	
C14.0	Pharynx, NOS
C14.2	Waldeyer's ring
C14.8	Overlapping lesion of lip, oral cavity and pharynx
Nasal cavity and middle ear	
C30.0	Nasal cavity (excludes Nose, NOS C76.0)
C30.1	Middle ear
Accessory sinuses	
C31.0	Maxillary sinus
C31.1	Ethmoid sinus
C31.2	Frontal sinus
C31.3	Sphenoid sinus
C31.8	Overlapping lesion of accessory sinuses
C31.9	Accessory sinus, NOS
Larynx	
C32.0	Glottis
C32.1	Supraglottis
C32.2	Subglottis
C32.3	Laryngeal cartilage
C32.8	Overlapping lesion of larynx
C32.9	Larynx, NOS
Thyroid gland*	
C73.9	Thyroid gland

Rarely SCC. Abbreviations: NOS, not otherwise specified.

2.4. How are head and neck cancers diagnosed?

2.4.1. Signs and symptoms

The symptoms of HNC vary according to the site of the tumour, but commonly include: a lump or ulcer in the mouth or throat that will not heal; difficulties when chewing, swallowing (dysphagia) or speaking (dysphonia); pain when swallowing (odynophagia); bleeding in the mouth; bad breath (halitosis); headache; tooth mobility without evidence of periodontal disease; persistent hoarseness or a change in voice; ear pain (otalgia) or stuffiness of ears; and swelling in the neck caused by an enlarged lymph node [19] [20]. Enlargement of a cervical lymph node (often painless) as the first presenting symptom is not unusual, especially in people with tumours in so-called "silent" sites e.g. base of tongue, supraglottis, and nasopharynx [3].

2.4.2. History and physical exam

The starting point of any diagnosis is the history and physical exam [21]. Here, the clinician will enquire about the presence and duration of any symptoms and record any other relevant medical, behavioural, and psycho-social history, paying particular attention to key risk factors such as prior cancer history, smoking history, alcohol consumption, occupational exposures, prolonged sun exposure and gastro-oesophageal reflux [22]. Following the history, the physician performs a comprehensive physical examination of the head and neck region, regardless of suspected primary site, owing to the frequent occurrence of multiple primary tumors in people with HNC. The examination follows a systematic approach, including: inspecting the face for asymmetry or swelling, examining the skin for ulcers, pigmentation or suspicious lesions, looking for abnormal discharge, bleeding, effusions from the ears or nose, and palpating the tongue and lymph nodes [22]. It is also helpful to listen to the individuals' voice and speech because this can provide an indication of the tumour location e.g. a raspy, hoarse voice may point towards a laryngeal neoplasm [23]. The doctor may also examine the nose and throat using a nasendoscope, a thin flexible tube that has a light and camera on the end [24]. Before inserting the nasendoscope, a local anesthetic is sprayed into the nostril to numb the area.

2.4.3. Diagnostic tests

To confirm a diagnosis of cancer and to determine whether it has spread, clinicians use a combination of biopsy and imaging tests. A biopsy is the removal of a small sample of tissue

to check for the presence of cancerous cells under the microscope. The common methods used in cases of suspected HNC are incision or punch biopsies and fine needle aspiration cytology (FNAC) [25]. The type of biopsy used will depend on the size and location of the tumour. FNAC is usually performed alongside an ultrasound. The ultrasound scanner uses high frequency sound waves to create a detailed image of the neck, which is then used to guide the needle to the correct position. A small sample of cells is then drawn up through the needle, like a blood test. A biopsy may also be carried out alongside a laryngoscopy. Similar to a nasendoscopy, laryngoscopy involves inserting a tube with a light and camera on the end into the mouth and throat, but it is usually done under a general anaesthetic [24]. If a diagnosis has not been made after extensive clinical evaluation and FNAC, an open excisional biopsy may be performed (i.e. the entire growth or lesion is removed).

The goals of imaging in HNC are: to determine the extent of the primary tumour and whether it has spread to regional lymph nodes; to detect whether the cancer has spread (metastasised) to other parts of the body; and to detect synchronous primary tumours. All these factors contribute substantially to prognosis and management of the disease. Staging of the tumour or node is upgraded from the original clinical stage in at least 30% of cases following imaging [3]. The imaging modality of choice is governed by the anatomic region under consideration [26]. For instance, tumours of the larynx, oropharynx, and hypopharynx are frequently imaged with computerised tomography (CT) because this technique is less affected by breathing and swallowing artefacts than other scanning procedures such as magnetic resonance imaging (MRI) [27]; It does however carry a radiation cost. The main advantage of MRI is that it is excellent for assessing the soft tissue extent of the tumour [28] i.e., it is able to distinguish between a mass and the surrounding soft-tissue structures. Other potential imaging techniques include nuclear magnetic resonance (NMR) scans, positron emission tomography (PET) scans, and chest X-rays, which are used to identify metastases or a second primary [26].

2.5. Staging of head and neck cancer

Clinical staging is paramount to the successful management of HNC and provides a common language through which physicians can communicate. It is the standard process of describing where the cancer is and how much of it there is. This information comes from various tests, including physical examination, imaging, and biopsies of affected areas. Knowing the cancer's stage helps clinicians develop a prognosis and determine the most appropriate course of treatment for their patient.

2.5.1. TNM staging

The tumour, node, metastasis (TNM) staging system is the most commonly used staging system across the world. It was developed by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC) [29]. The system is based on local tumour growth (T), the degree of regional lymph node involvement (N), and the absence or presence of distant metastases (M). The most common site of metastases is the lungs, followed by the liver and bones [30].

Two types of TNM staging are applied in practice: i) clinical TNM (cTNM) staging, which is based on clinical examination in addition to ancillary techniques such as radiological investigations (i.e. MRI or CT); and ii) pathological TNM (pTNM) staging, which is derived from the histopathological examination of the tumour specimen (i.e. surgical pathology reports) [31]. The pTNM is considered superior to cTNM in that it provides better prediction of prognosis [32]. Thus, it is generally used when deciding if adjuvant therapy is needed. Previous work suggests a concordance rate between clinical and pathological N classification of between 43% and 83% depending on the tumour site [33].

As a general rule, the T stage will take a value between 0-4 depending on the size and extent of the tumour; the N stage will be a number between 0–3 depending on the number of local lymph nodes involved, their location and their size, and the M stage will be either a 0 if the cancer is restricted to its primary site, or a 1 if it has spread to other parts of the body [34]. Specific subdivisions may exist for each stage. For instance, T4a disease is considered “moderately advanced” but resectable, whilst T4b disease is deemed “very advanced” and unresectable, due to the aggressive nature of the cancer [35]. The staging of the different types of HNC are all slightly different, although most HNC sites (except the thyroid and nasopharynx) use the same classification system for regional lymph nodes [36]. Tables 2a and 2b provide the T-and N- classification according to anatomical site according to the 8th edition of the AJCC Cancer staging manual.

Table 2a: Tumour (T) staging according to anatomical site [37].

Oral cavity	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
Tis	Carcinoma *in situ.*
T1	Tumour <2 cm in greatest dimension.
T2	Tumour > 2 but < 4 cm in greatest dimension.
T3	Tumour > 4 cm.
T4a	Moderately advanced local disease. Tumour invades through cortical bone, inferior alveolar nerve, floor of mouth, or skin of face—that is, chin or nose. Tumour invades adjacent structures (e.g., through cortical bone, into deep [extrinsic] muscle of tongue [genioglossus, hypoglossus, palataglossus, and styloglossus], maxillary sinus, skin of face).
T4b	Very advanced local disease. Tumour invades masticator space, pterygoid plates, or skull base and/or encases internal carotid artery.
Oropharynx	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
Tis	Carcinoma *in situ.*
T1	Tumour <2 cm in greatest dimension.
T2	Tumour > 2 cm but < 4 cm in greatest dimension.
T3	Tumour > 4 cm or extension to lingual surface of epiglottis.
T4a	Moderately advanced local disease. Tumour Invades the larynx, deep/extrinsic muscle of tongue, medial pterygoid, hard palate, or mandible.
T4b	Very advanced local disease. Tumour invades prevertebral space, encases carotid artery, or invades mediastinal structures.
Larynx	
TX	Primary tumour cannot be assessed
T0	No evidence of primary tumour
Tis	Carcinoma *in situ*
Supraglottis	
T1	Tumour limited to one sub site, with normal vocal cord mobility
T2	Tumour invades mucosa of more than one of the supraglottis or glottis or region outside the supraglottis (e.g., mucosa of base of tongue, vallecula, medial wall of pyriform sinus) without fixation of the larynx
T3	Tumour limited to the larynx with vocal fold fixation and/or invades any of the following: postcricoid area, pre-epiglottic tissues, paraglottic space, and/or inner cortex of thyroid cartilage
T4a	Moderately advanced local disease. Tumour invades through the thyroid cartilage and/or invades tissues beyond the larynx (e.g., trachea, soft tissues of neck including deep extrinsic muscle of the tongue, strap muscles, thyroid, or oesophagus).
T4b	Invades prevertebral space, encases carotid artery, or invades mediastinal structures.

Table 2a continued.

Glottis	
T1	Tumour limited to vocal cord(s), with normal mobility.
T1a	Tumour limited to one vocal cord.
T1b	Tumour involves both vocal cords.
T2	Tumour extends to supraglottis, and/or subglottis, and/or with impaired vocal cord mobility.
T3	Tumour limited to the larynx with vocal fold fixation and/or invasion of paraglottic space, and/or inner cortex of the thyroid cartilage.
T4a	Moderately advanced local disease. Tumour invades the outer cortex of the thyroid cartilage and/or invades tissues beyond the larynx (e.g., trachea,
T4b	Very advanced local disease. Tumour invades prevertebral space, encases carotid artery, or invades mediastinal structures.
Subglottis	
T1	Tumour limited to subglottis.
T2	Tumour extends to the vocal cord(s) with normal/impaired mobility.
T3	Tumour limited to the larynx with vocal cord fixation.
T4a	Moderately advanced local disease. Tumour invades cricoid or thyroid cartilage and/or invades tissues beyond the larynx (e.g., trachea, soft tissues of the neck including deep extrinsic muscles of the tongue, strap muscles, thyroid, or oesophagus).
T4b	Very advanced local disease. Tumour invades prevertebral space, encases carotid artery, or invades mediastinal structures.
Salivary gland	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
T1	Tumour <2 cm in greatest dimension without extraparenchymal Extension.
T2	Tumour >2 cm but <4 cm in greatest dimension without extraparenchymal extension**
T3	Tumour >4 cm and/or tumour having extraparenchymal extension**.
T4a	Moderately advanced local disease. Tumour invades the skin, mandible, ear canal, and/or facial nerve.
T4b	Very advanced local disease. Tumour invades the skull base and/or pterygoid plates and/or encases the carotid artery.
Hypopharynx	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
Tis	Carcinoma *in situ.*
T1	Tumour limited to one subsite of the hypopharynx and is <2 cm in greatest dimension.
T2	Tumour invades more than one subsite of the hypopharynx or an adjacent site, or measures >2 cm but < 4 cm in greatest dimension without fixation of the hemilarynx or extension to the oesophagus.
T3	Tumour >4 cm in greatest dimension or with fixation of the hemilarynx or extension to the oesophagus.
T4a	Moderately advanced local disease. Tumour invades thyroid/cricoid cartilage, hyoid bone, thyroid gland, oesophagus, or central compartment soft tissue*

Table 2a continued.

T4b	Very advanced local disease. Tumour invades prevertebral fascia, encases carotid artery, or involves mediastinal structures.
Nasal cavity and paranasal sinuses	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
Tis	Carcinoma *in situ.*
maxillary sinus	
T1	Tumour limited to the maxillary sinus mucosa with no erosion or destruction of bone.
T2	Tumour causing bone erosion or destruction, including extension into the hard palate and/or middle nasal meatus, except extension to the posterior wall of the maxillary sinus and pterygoid plates.
T3	Tumour invades any of the following: bone of the posterior wall of the maxillary sinus, subcutaneous tissues, floor or medial wall of the orbit, pterygoid fossa, or ethmoid sinuses.
T4a	Moderately advanced local disease. Tumour invades anterior orbital contents, skin of cheek, pterygoid plates, infratemporal fossa, cribriform plate, sphenoid or frontal sinuses.
T4b	Moderately advanced local disease. Tumour invades any of the following: orbital apex, dura, brain, middle cranial fossa, cranial nerves other than maxillary division of trigeminal nerve [V_2], nasopharynx, or clivus.
nasal cavity & Etmoid sinus	
T1	Tumour restricted to any one subsite, with or without bony invasion.
T2	Tumour invades two subsites in a single region or extending to involve an adjacent region within the nasoethmoidal complex, with or without bony invasion.
T3	Tumour extends to invade the medial wall or floor of the orbit, maxillary sinus, palate, or cribriform plate.
T4a	Moderately advanced local disease. Tumour invades any of the following: anterior orbital contents, skin of nose or cheek, minimal extension to anterior cranial fossa, pterygoid plates, sphenoid or frontal sinuses.
T4b	Very advanced local disease. Tumour invades any of the following: orbital apex, dura, brain, middle cranial fossa, cranial nerves other than V_2, nasopharynx, or clivus.
Nasopharynx	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
Tis	Carcinoma *in situ*.
T1	Tumour confined to the nasopharynx or tumour extends to the oropharynx and/or nasal cavity without parapharyngeal extension.
T2	Tumour with parapharygeal extension.
T3	Tumour involves bony structures of skull base and/or paranasal sinuses.
T4a	Tumour with intracranial extension and/or involvement of cranial nerves, hypopharynx, orbit, or with extension to the infratemporal fossa/ masticator space.

Table 2a continued.

Thyroid gland	
TX	Primary tumour cannot be assessed.
T0	No evidence of primary tumour.
T1	Tumour <2 cm or less in greatest dimension, limited to the thyroid.
T1a	Tumour <1 cm or less, limited to the thyroid.
T1b	Tumour >1 cm but <2 cm in greatest dimension, limited to the thyroid.
T2	Tumour > 2 cm but <4 cm in greatest dimension, limited to the thyroid.
T3	Tumour >4 cm in greatest dimension, limited to the thyroid or any tumour with minimal extrathyroid extension (e.g., extension to sternothyroid muscle or perithyroid soft tissues).
T4a	Moderately advanced local disease. Tumour of any size extending beyond the thyroid capsule to invade subcutaneous soft tissues, larynx, trachea, oesophagus, or recurrent laryngeal nerve.
T4b	Very advanced local disease. Tumour invades prevertebral fascia or encases the carotid artery or mediastinal vessels.

Table 2b: Classification of lymph nodes by anatomical site [37]

Excluding the nasopharynx and thyroid	
NX	Regional lymph nodes cannot be assessed.
N0	There is no regional nodes metastasis.
N1*	Metastasis is in a single ipsilateral lymph node, ≤ 3 cm in greatest dimension.
N2*	Metastasis is in a single ipsilateral lymph node, > 3 cm but < 6 cm in greatest dimension; or metastasis is in multiple ipsilateral lymph nodes, none > 6 cm in greatest dimension; or metastasis is in bilateral or contralateral lymph nodes, none > 6 cm in greatest dimension.
N2a*	Metastasis is in a single ipsilateral lymph node, > 3 cm but ≤ 6 cm in greatest dimension.
N2b*	Metastasis is in multiple ipsilateral lymph nodes, none > 6 cm in greatest dimension.
N2c*	Metastasis is in bilateral or contralateral lymph nodes, none > 6 cm in greatest dimension.
N3*	Metastasis is in a lymph node > 6 cm in greatest dimension.
Nasopharynx	
NX	Regional lymph nodes cannot be assessed.
N0	No regional lymph node metastasis.
N1	Unilateral metastasis in cervical lymph node(s), 6 cm or less in greatest dimension, above the supraclavicular fossa, and/or unilateral or bilateral retropharyngeal lymph nodes, 6 cm or less in greatest dimension.
N2	Bilateral metastasis in cervical lymph node(s), 6 cm or less in greatest dimension, above the supraclavicular fossa.
N3	Metastasis in lymph node >6 cm and/or to supraclavicular fossa.
N3a	Greater than 6 cm in dimension.
N3b	Extension to the supraclavicular fossa.

Table 2b. continued.

Thyroid	
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis.
N1	Regional lymph node metastasis.
N1a	Metastasis to Level VI (pretracheal, paratracheal, and prelaryngeal/ Delphian lymph nodes).
N1b	Metastasis to unilateral, bilateral, or contralateral cervical Levels I, II, III, IV, or V) or superior mediastinal lymph nodes (Level VII).

2.5.2. Prognostic staging

Based on the TNM numbers, the cancer is assigned an overall stage of 0 to IV. In most cases, early-stage disease is denoted as stage I or stage II and advanced stage disease is denoted as stage III or stage IV. Staging is slightly different for each particular type of HNC, but can be generalised as follows [38]:

- *Stage 0 (also called carcinomas in situ)*: cancers are only found in the squamous cells where they began;
- *Stage I:* cancers have grown into the tissue (up to 2cm across), but they have not yet invaded nearby lymph nodes;
- *Stage II:* cancers have extended into nearby structures (2-4 cm across), but have still not spread to the lymph nodes or other parts of the body;
- *Stage III:* cancers may have extended further (more than 4 cm across), or they may have spread to a single ipsilateral lymph node (a node located on the same side of the head or neck as the primary tumour);
- *Stage IVa:* cancers may be any size with more than one ipsilateral lymph node involved, or they may have extended into the major structures of the head and neck (e.g. the tumour has invaded the skull base or encases the carotid artery), in which case they may or may not have spread to one ipsilateral lymph node that is smaller than 3cm across;
- *Stage IVb:* cancers have not spread to other sites in the body, but one of the following is true: they have grown into major structures of the head and have spread to one or multiple ipsilateral lymph nodes that are more than 6 cm across; or, they are any size with bilateral or contralateral lymph node metastases.
- *Stage IVc:* cancers may be any size but have metastasised to distant sites of the body.

2.5.3. Pros and cons of the TNM system

The TNM classification system plays a critical role in cancer care and research. It assists in making and communicating diagnoses, formulating prognoses and defining or stratifying study groups. Its main advantages are that it is widely applicable, easy to adhere to and crucially, it is universally accepted [39]. Important limitations do exist, however. A major limitation of the present TNM classification system for HNC is that it is based almost exclusively on tumour morphology, with the exception of OPC which will be discussed below. Because individual-based prognostic factors (e.g. age, gender and comorbidity) and biological and molecular markers are not taken into account, this limits its ability to provide individualised prognostication [39].

As alluded to above, OPC has its own distinct TNM staging system. It was first introduced in the eighth edition of the AJCC Staging Manual. The new staging schema aims to provide improved discrimination and predictive ability and is based on emerging knowledge of the different clinical behaviours of human papilloma virus (HPV)-associated and non-HPV-associated oropharyngeal tumours. Using the new prognostic stage groups for HPV-associated OPC (Table 3), people who would have been considered stage IV under the old staging system may now be classified as stage I, II or III. The stratification of OPC into HPV-associated and non-HPV associated types is dependent on immunohistochemical staining of the tumour suppressor protein p16 (cyclin-dependent kinase 2A), which is highly correlated with HPV infection in HNC [40]. Further details of p16 will be provided in the next chapter.

Table 3: AJCC (8th edition) prognostic stage groups for HPV-associated (p16+) OPC.

	N category	M category	Stage group
Clinical			
T0, T1 or T2	N0 or N1	M0	I
T0, T1 or T2	N2	M0	II
T3	N0, N1, or N2	M0	II
T0, T1, T2, T3 or T4	N3	M0	III
T4	N0, N1, N2 or N3	M0	III
Any T	Any N	M1	IV
Pathological			
T0, T1 or T2	N0 or N1	M0	I
T0, T1 or T2	N2	M0	II
T3 or T4	No or N1	M0	II
T3 or T4	N2	M0	III
Any T	Any N	M1	VI

2.6. Head and neck cancer treatment

The treatment of HNC is dictated by the location of the tumour, the stage of the disease, and the individual's age and general health. Eliminating the cancer is always the primary goal, however, preserving the function of nearby organs, nerves and tissues is also important as this will affect the person's quality of life (QoL) [41]. For this reason, treatment decisions are made by a multidisciplinary team of surgeons, medical oncologists, radiation oncologists and rehabilitation specialists [42]. The main treatment options available are surgery, radiation therapy, chemotherapy, targeted therapy, or a combination of treatments.

2.6.1. Treating early stage cancers

If the cancer is small and has not spread to lymph nodes or metastasised elsewhere, it can usually be treated with single-modality surgery or radiotherapy [43]. The choice between surgery or radiation is influenced by the location of the primary tumour and the likely functional sequelae. For instance, surgery is typically preferred for floor of the mouth cancers due to the risk of complications from radiotherapy, but base of tongue tumours are less amenable to surgical resection, which could cause permanent deficits in speech and swallowing. In circumstances such as this, where surgery would be associated with

unacceptable morbidity, radiotherapy would be preferred. Sometimes, radiotherapy may be used in combination with surgery to improve prognosis and reduce the risk of locoregional recurrence. It may be given before surgery (neoadjuvant radiotherapy), to try to shrink the tumour so that it is easier to resect, or it may be given after surgery if the cancer is very large or it is not feasible to remove it completely [44]. Adjuvant radiotherapy should ideally begin 4–6 weeks following primary surgery [38].

The most common type of radiation therapy is called external-beam radiation therapy (EBRT), which delivers radiation from a machine outside the body [45]. An EBRT regime, or schedule, is usually fractioned, which means that the total dose of radiation to be given is divided into small doses (fractions) that are administered over a prolonged period. The radiation dose depends on tumour size; however, for early stage disease, conventional radiation schedules deliver a single fraction of 2 Gy/day, 5 days a week for 7 weeks [46].

Intensity-modulated radiation therapy (IMRT) describes an advanced type of EBRT that shapes the radiation beam to fit precisely around the tumour [47]. This means that the cancer cells receive a high dose of radiation, whilst the surrounding healthy cells receive a lower dose, reducing the risk of long-term side-effects. In one UK-based randomised trial comparing conventional radiotherapy (control) with IMRT, IMRT was shown to reduce radiation-induced xerostomia i.e. dry-mouth (the most common long-term side effect of standard radiotherapy), from 75% to 39% twelve months post-treatment [48]. A further type of radiation therapy used to treat HNC is Interstitial/intracavitary brachytherapy, which delivers radiation directly to the tumour using removable implants (radioactive needles, wires or seeds) [42]. Again, this approach can spare nearby critical structures such as the brain, spinal cord and eyes, from excessive radiation doses. Brachytherapy is often used in combination with externally delivered radiotherapy, or to treat recurrent cancers.

2.6.2. Treating locally advanced cancer

People with locally advanced tumours of the head and neck (stages III-IV, excluding T4b tumours) are usually treated by surgery combined with post-operative chemoradiation. As a single modality, chemotherapy has not been found to be effective in curing HNC, but when administered concurrently with radiotherapy, chemotherapy has been found to improve locoregional control and survival, with an absolute survival benefit of around 6.5% at five years (compared to radiotherapy alone) [49]. Chemotherapy also plays an important role in the palliative treatment of HNC, to relieve symptoms and improve QoL. The most commonly used chemotherapeutic agents are cisplatin (cis-diamminedichloroplatinum), and carboplatin

(a derivative of cisplatin), which have similar modes of action. They crosslink with deoxynucleic acid (DNA), inhibiting DNA repair mechanisms, which ultimately results in apoptosis, or cell death.

When treating recurrent or metastatic cancers, platinum-based chemotherapy is often augmented with a drug called Cetuximab (also known by its brand name Erbitux) [50]. Cetuximab is a recombinant monoclonal antibody (mAB) that binds with high affinity to the human epidermal growth factor receptor (EGFR), which is overexpressed in more than 90% of HNCs [51]. Binding inhibits the proliferation of cancer cells, which depend on EGFR activation for growth [52]. Cetuximab is often referred to as a "targeted drug", in that it is directed specifically towards molecules that promote the proliferation of cancer cells, in contrast to standard chemotherapy or radiotherapy, which act systemically. The treatment efficiency of Cetuximab is relatively low however, around 36% in combination with chemotherapy.

2.7. Summary

HNC can occur at a large number of subsites including the lip, mouth and pharynx. The presenting symptoms vary depending on the site of the primary tumour but include hoarseness, difficulties swallowing and pain in the middle ear. The diagnosis is usually confirmed by biopsy of the primary site and FNA of any enlarged lymph nodes. If inconclusive, an open excisional biopsy may be performed. Imaging is crucial for assessing the site and extent of the cancer and contributes substantially to treatment decisions and prognosis. The most commonly used modalities are CT and MRI scans. MRI is particularly useful for assessing the relationship of cancer boundaries to normal anatomical structures and, unlike CT, it does not use damaging ionizing radiation; however, artefacts from swallowing, breathing and coughing can limit the quality of MRI. HNC staging is done according to the AJCC and UICC classification system, an anatomically-based system based on the extent of the tumour (T), the extent of spread to the lymph nodes (N), and the presence or absence of metastasis (M). Small, early-stage tumours are typically treated with radiotherapy or surgery, whilst larger, more advanced cancers are treated with surgery and post-operative chemotherapy, though the particular location of the tumour and the likely long-term sequelae usually determine the decision on management. Inoperable cancers may be treated with combinations of chemo- and radiotherapy, but in some situations only the symptoms of disease can be treated. The distribution, patterns and determinants of HNC are discussed in the next chapter.

Chapter 3: Head and neck cancer epidemiology, biomarkers and prognosis

3.1. Introduction

Chapter two provided a general introduction to HNC, a diverse group of tumours that arise in epithelial lining of the upper digestive tract. The ways in which these cancers are diagnosed and treated was described and the main risk factors for development were briefly mentioned. This chapter reviews the epidemiology of the disease, and reports on previously studied epigenetic and metabolomic biomarkers and other prognostic factors. It begins with an overview of HNC prevalence, incidence and mortality, focusing mainly on the most common HNC sub-sites. A more in-depth discussion of the major risk factors is then provided. The final section of this chapter focusses on factors that have been purported to influence HNC outcomes and which may provide prognostic value, including epigenetic and metabolomic biomarkers, and lifestyle factors.

3.2. Prevalence of head and neck cancer

HNC is the sixth most common malignancy in the world today, with more than 550,000 cases reported annually [53]. The actual prevalence of the disease, i.e. the number of people living with HNC, is difficult to determine because cancer registries are not compulsory in many developing countries where the burden of disease is high e.g. India [54-59]. For this reason, It has been suggested that what appears in the literature is only the 'tip of the iceberg' [58]. In the UK alone, where HNC accounts for 3% of all cancers (compared to an estimated 30% of cancers in India [60]), there were 62,500 people living with a diagnosis of HNC in 2015. More than half of these people were aged 65 years or older [61]. The majority of people living with HNC in the UK have lip and oral cavity cancers, with an estimated 5-year prevalence rate of 31.1 per 100,000 (Table 4). Oropharyngeal and laryngeal cancer are the next most common HNCs with 5-year prevalence rates of 20.8 and 12.3 per 100,000 persons respectively [62]. A similar pattern of prevalence is seen across Europe and the USA e.g. the corresponding rates for oral cavity, oropharyngeal and laryngeal cancer in the USA are 27.0, 17.4 and 16.7.

Table 4: Estimated number of prevalent cases of HNC in the UK (2018) as a proportion, ages 20-85+ years.

	Overall			Males			Females		
HNC site	**1-year**	**3-year**	**5-year**	**1-year**	**3-year**	**5-year**	**1-year**	**3-year**	**5-year**
Lip, oral cavity	8.5	21.2	31.1	11.5	28.6	42.0	5.6	14.2	20.9
Oropharynx	5.2	13.8	20.8	8.1	21.1	31.8	2.4	6.8	10.2
Larynx	3.3	8.3	12.3	5.5	13.9	20.6	1.2	3.1	4.7
Salivary gland	1.0	2.4	3.3	1.1	2.7	3.8	0.9	2.1	3.0
Hypopharynx	0.9	1.8	2.3	1.5	2.9	3.6	0.4	0.8	1.0
Nasopharynx	0.4	1.0	1.6	0.6	1.6	2.4	0.2	0.5	0.8

Proportions per 100,000. Figures based on GLOBOCAN 2018 projections [62].

3.3. HNC Incidence rates

Cancer incidence refers to the number of new cases of cancer reported in the population per unit time. It is usually expressed as the number of cases per 100,000 person years at risk [63]. This is termed the crude incidence rate. When examining cancer trends, crude incidence (and prevalence) rates can be misleading if the populations or geographic regions under consideration have different age profiles. For instance, if two equally sized countries - *A* and *B*, report identical numbers of cancer but country *A* has a much younger age structure than *B*, then the incidence rate in *A* would in effect be higher because the likelihood of developing cancer increases with age [64]. Consequently, incidence rates are often calculated as age-standardised rates (ASR) [65] [66] to allow direct comparisons.

3.3.1. UK incidence rates

In the UK, there are around 13,500 new cases of HNC diagnosed each year [62], making it the eighth most commonly diagnosed cancer [67]. Consistent with the prevalence, most of these cases (80-90%) occur in the oral cavity, oropharynx and larynx [64]. The number of incident cases is much higher in males than in females, with over two-thirds (70%) of all cases occurring in men [64]. The exact ratios of males to females differs by anatomical subsite (Table 5). Laryngeal cancer has a particularly high incidence among males, with a male: female ratio of around 3:1.

Table 5: UK estimated age-standardised incidence rates of head and neck cancer by site and gender, all ages (2018).

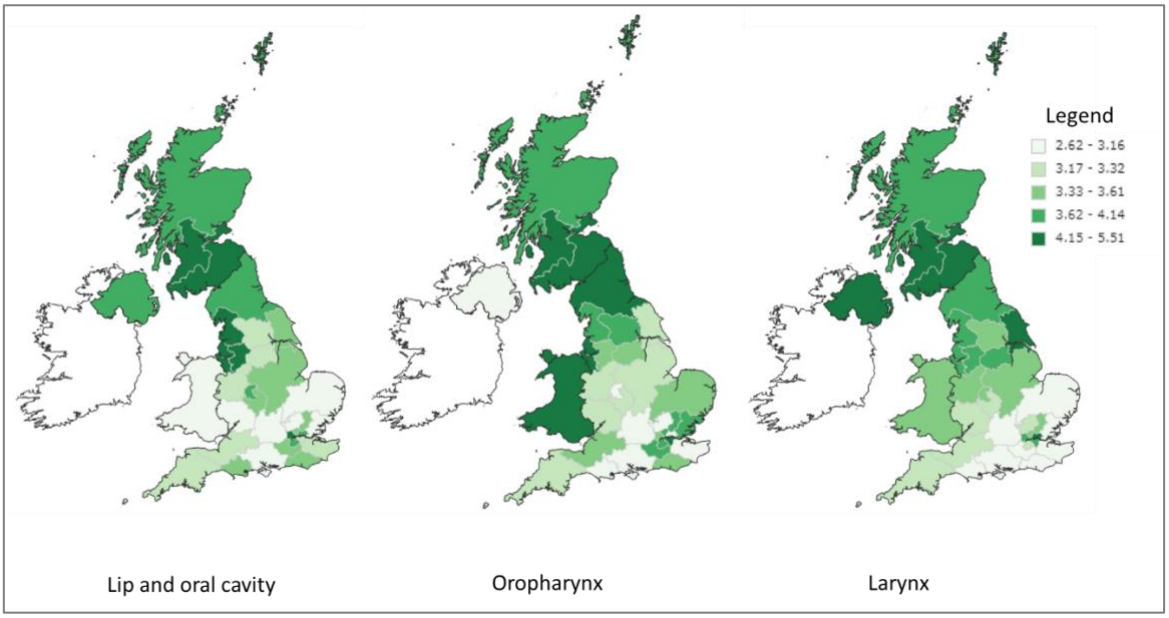
		Numbers			ASIR		
ICD	**Cancer**	**All**	**Males**	**Females**	**All**	**Males**	**Females**
C00-06	Lip, oral cavity	6 087	3 880	2 207	5.0	6.8	3.3
C32	Larynx	2 482	2 010	472	1.9	3.2	0.7
C11	Nasopharynx	269	199	70	0.3	0.4	0.1
C09-10	Oropharynx	3 049	2 313	736	2.9	4.6	1.3
C12-13	Hypopharynx	782	612	170	0.6	1.0	0.2
C07-08	Salivary gland	803	482	321	0.7	0.8	0.6
	Total	13,472	9,496	3,976			

Figures based on GLOBOCAN 2018 projections [62]. Abbreviations: ICD, international classification of diseases; ASIR, age-standardised rates per 100,000, calculated using the direct method and the world standard population [68].

There are notable regional differences in HNC incidence, as illustrated in Figure 7. Northern England has higher incidence rates in both males and females compared to the rest of England [69-71], whilst Scotland has higher incidences than the rest of the UK. In 2015, the European ASIR for HNC in Scotland was 24.6 (37.0 for males and 13.8 for females), compared to 19.2 for England (28.1 for males and 11.2 for females) [67].

Incidence patterns of HNC likely reflect differences in the use of tobacco and alcohol, two of the major established risk factors for HNC. In England, an estimated 15.8% of the UK adult population smoke tobacco [72] and 21% drink excessive amounts of alcohol (more than 14 units a week) [73], but the prevalence of these high-risk behaviours is not distributed equally [74-77]. A north–south divide, which is explained to some extent by socioeconomic inequalities [78 79], has been reported for smoking, with higher rates observed in northern regions of the country [77 80]. Similarly, in Scotland, the vast majority of people developing HNC come from the most disadvantaged areas where the prevalence of tobacco use is high [80-82]. For alcohol drinking, a more complex pattern exists, with lower rates of excessive consumption in central and eastern regions of England and an east versus west divide in the prevalence of alcohol dependency [81 82].

Figure 7: Geographical distribution of oral cavity, oropharyngeal and laryngeal cancer incidences, UK (2007-2009).

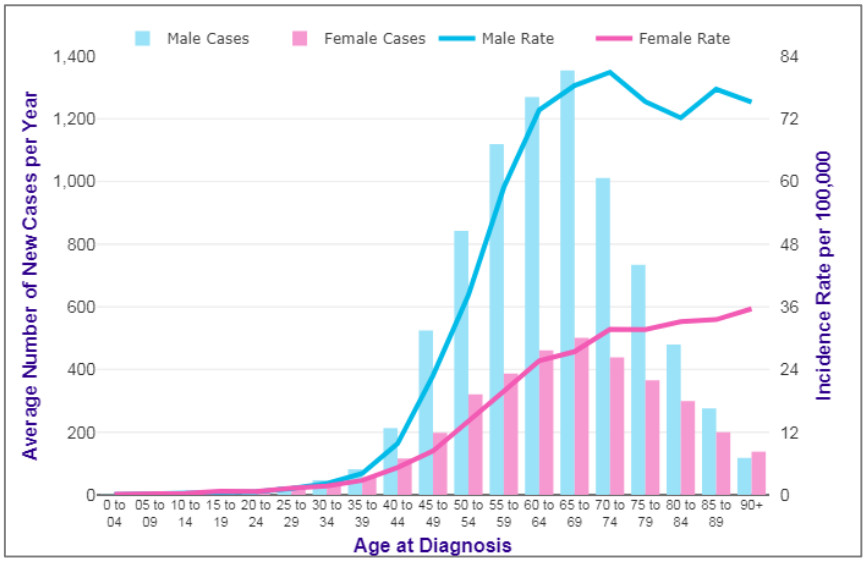


Incidence data are displayed using directly age-standardised rates. Rates are presented per 100,000 population, using European population as standard. Credit: National Cancer Information Service (NCIS) [83].

As stated previously, HNC is strongly related to age. In UK males, there is a sharp upward trend in incidence rates (all sites combined) between the ages of 35-39 years, after which point rates increase more steadily, eventually peaking at around 70-74 years (Figure 8). In females, there is a much more gradual rise in incidence rates from the age of 35-39 years, with most HNC cases occurring in the 90+ age group [64] (Figure 8). Overall, for both sexes combined, more than a fifth (22%) of all cases in the UK are diagnosed in people aged 75 years or older [64].

The number of cancer registrations reported across different age groups differs between subsites. According to the Office for National Statistics (ONS), most laryngeal cancer registrations in England occur between the ages of 70-74 years in males and 65-74 years in females; for lip, oral cavity and pharyngeal cancer, the majority of new cases for both males and females occur between the ages of 65-69 years; and for oropharyngeal cancer, most registrations occur between the ages of 60-64 years in males and 70-74 years in females [71].

Figure 8: Average number of incident cases of HNC per year and age-standardised incidence rates per 100,000 persons in the UK (2013-2015).



Data is for ICD-10 C00-C14, C30-C32. Credit: CRUK [64]

3.3.2. Global variations in incidence rates

Table 6 presents the estimated global age-standardised incidence rates (ASIR) of HNC by site and gender based on the world population. As will be demonstrated below, however, there is substantial variation in the anatomic distribution of HNC, both between countries and within them and this is largely driven by differences in the distribution of risk factors.

A full description of HNC incidence rates by site, country and region is outside the bounds of this thesis but the next section provides an overview of some of the major global patterns. An emphasis is placed on the most common anatomical sites (i.e. oral cavity, oropharynx and larynx) but other HNCs will be discussed in brief. Where the term “oral cancer” is used, ICD codes will be provided in brackets since there is inconsistency in the literature regarding the definition of this term. Unless otherwise stated, ASIRs are derived from GLOBOCAN 2018 projections [62] and represent the ASIR per 100,000 world population (all ages).

Table 6: Global estimated age-standardized incidence rates of head and neck cancer by site and gender, all ages (2018).

	Number			ASIR		
Cancer site	**All**	**Males**	**Females**	**All**	**Males**	**Females**
Lip, oral cavity	354 864	246 420	108 444	4.0	5.8	2.3
Larynx	177 422	154 977	22 445	2.0	3.6	0.5
Nasopharynx	129 079	93 416	35 663	1.5	2.2	0.8
Oropharynx	92 887	74 472	18 415	1.1	1.8	0.4
Hypopharynx	80 608	67 496	13 112	0.9	1.6	0.3
Salivary gland	52 799	29 256	23 543	0.6	0.7	0.5
Total	887,659	666,037	221,622			

*Figures based on GLOBOCAN 2018 projections [62]. Abbreviations: **ICD**, international classification of diseases; **ASIR**, age-standardised incidence rates per 100,000, calculated using the direct method and the world standard population [68].*

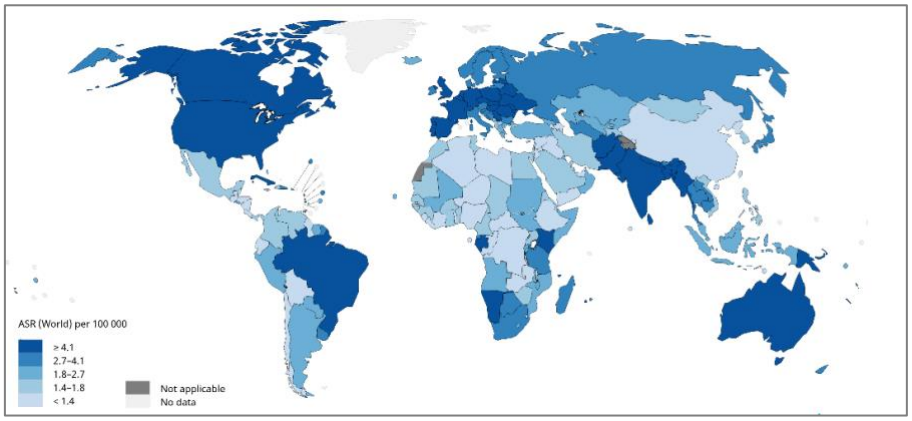
3.3.2.1. Lip and oral cavity cancer

As illustrated in Figure 9 (a.), a particularly high incidence of lip and oral cavity cancer can be seen in South and Southeast Asia (e.g. India, Pakistan and Bangladesh). Among Indian males specifically, ASIRs of lip and oral cavity cancer surpass those for any other cancer (AISR of 13.9), whilst in Indian females they ranked as the third most commonly diagnosed malignancies after breast and cervical cancer, with an ASIR rate of 4.3 [62].

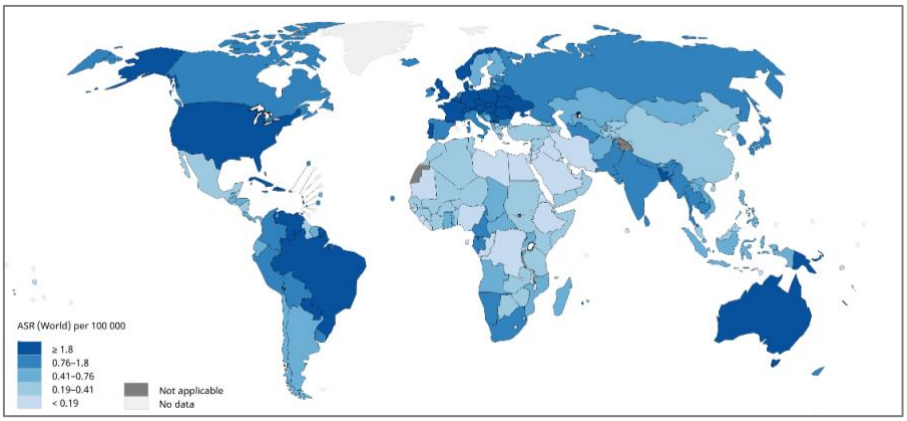
The aetiology of lip and oral cavity cancer in Southern Asia is dominated by tobacco use, especially smokeless tobacco use [84]. Of the 42.4% of men and 14.2% of women in India who use tobacco, 29.6% and 12.8% use smokeless tobacco, respectively [85]. Smokeless tobacco products such as betel quid/areca nut, gutkha and mishri, are often held in the mouth for extended periods of time causing prolonged exposure to carcinogenic tobacco-specific nitrosamines [86 87]. Their use is most common among people living in rural and low-income communities-both male and female [88].

Figure 9: Global incidence of HNC, by sub-site.

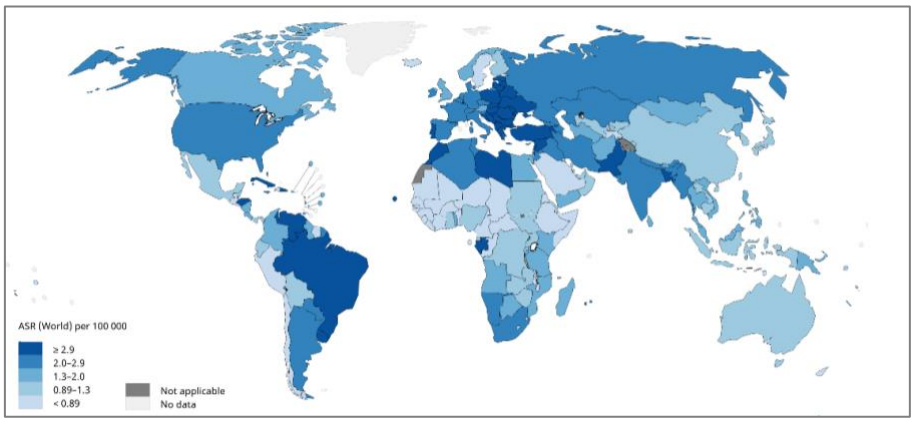
a) lip and oral cavity



b) oropharynx



c) Larynx



All of the above maps present the estimated ASIRs in both sexes, all ages. Credit: GLOBOCAN 2018 Graph production: IARC (http://gco.iarc.fr/today) World Health Organization.

Alcohol drinking is also a risk factor for oral cavity cancer in this region, although it plays less of a significant role than tobacco use. In a meta-analysis looking at the magnitude of the effect of tobacco smoking, betel quid chewing and alcohol drinking on oral cancer (ICD-10, codes C00–C06, C09, C10) in Southeast Asia (n=59,000), the pooled odds ratio (OR) for alcohol drinkers was 2.2 (95% CI: 1.6 to 3.0), compared to 3.6 for smokers (95% CI: 1.9 to 7.0) and 7.9 (95% CI: 6.7 to 9.3) for smokeless tobacco users. Among smoking-drinking-chewing subjects, the odds ratio increased to 40.1 (95% CI: 35.06 to 45.83) [89], suggesting an interaction effect.

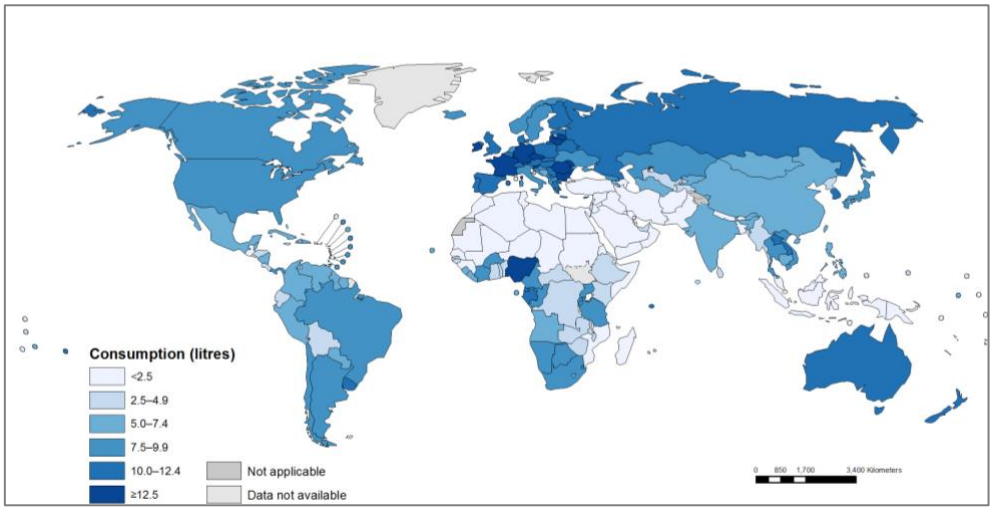
The prevalence of alcohol drinking in Southern Asia is relatively low overall [90] (Figure 10), largely due to religious and cultural practices. In many rural areas however, alcohol consumption rates are high [91] and therefore the proportions of oral cavity cancers attributed to alcohol use are expected to be greater. Low fruit and vegetable intakes and poor oral hygiene have also been suggested as important risk factors for oral cavity cancer among these communities [92 93].

Outside South and Southeast Asia, Europe has some of the highest rates of lip and oral cavity cancer. Of all the European countries, Hungary has the highest incidence rate overall (7.5), with high rates also noted in Latvia (6.9) and France (6.3). As illustrated in Figures 10 and 11, the prevalence of alcohol drinking and tobacco smoking in Europe are also high [94 95]. In Hungary for instance, 32% of men and almost 25% of women smoke; the rate among Latvian men is even higher at 48.9%. These figures are above the average for high human Development Index (HDI) countries [96] and probably explain, to some extent, the high incidence rates of oral cavity cancer in these countries.

In Hungary and in a few other countries of eastern Europe (e.g. Bulgaria, Poland and Romania), the high incidence of lip and oral cavity cancer have been linked both to the pattern of alcohol consumption and the type and quality of alcohol consumed [97]. Specifically, heavy episodic drinking or “binge drinking” (defined as drinking 60 grams or more of pure alcohol on at least one single occasion at least once per month [94]), and the consumption of home-produced wines and spirits is very common. Homemade alcoholic beverages often contain high levels of acetaldehyde [97] which, as will be discussed later, is an established human carcinogen.

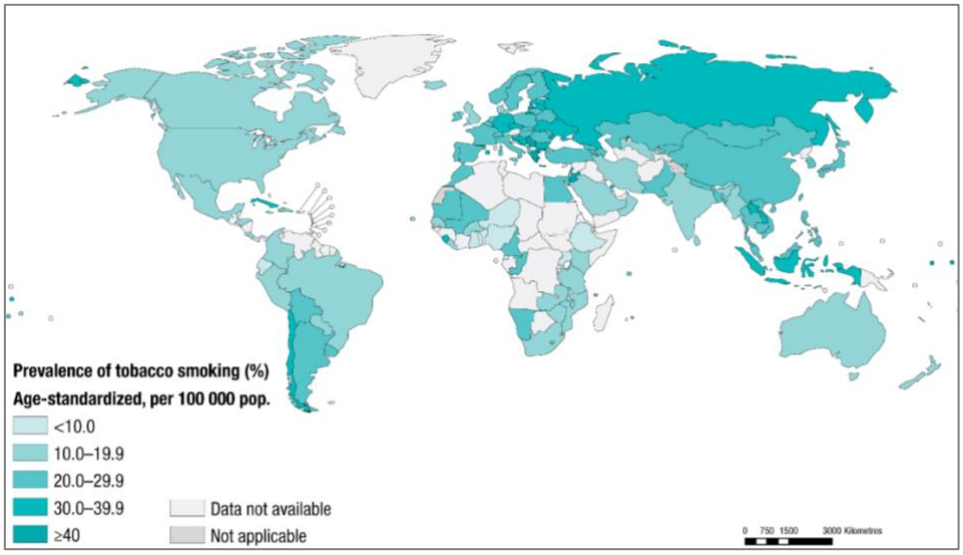
Australia also has a high incidence of lip an oral cavity cancer. In addition to tobacco and alcohol exposure, solar radiation (especially to the lips), is a significant risk factor for disease in this region [55].

Figure 10: Total alcohol per capita consumption (15+ years; in litres of pure alcohol), 2016.



Credit: WHO [98]

Figure 11: Global prevalence of tobacco smoking among persons aged 15 years and older (2015).



Age-standardised prevalence of tobacco smoking (%) per 100,000.
Credit: WHO [99]

3.3.2.2. Oropharyngeal cancer

Regions with the highest rates of OPC include Europe, Northern America, parts of South America (e.g. Paraguay, Brazil and Venezuela) and Australia (Figure 9b.). Within Europe, Hungary, Denmark and France have the highest ASIRs overall at 4.7, 4.5 and 4.4, respectively. This compares to the UK incidence rate of 2.9. Rates in the USA and Australia are broadly comparable to those in the UK (USA, 2.4; Australia 2.5). In all of these countries, the incidence of OPC is much higher in males than in females. In the USA for example, the ASIR for males is 4.1 compared to 0.8 in females whilst in Hungary and Brazil, the corresponding rates are 7.3 and 2.4 and 3.3 and 0.6.

Consistent evidence suggests that two main types of OPC exist; the first, which is common in developing countries, is driven by tobacco and alcohol use; the second, which is common in developed countries (e.g. Northern America and Europe), is driven by HPV infection [100] [101]. In Europe, the estimated fraction of OPCs attributable to HPV ranges from 24% (95% CI: 17 to 30) in Southern Europe to 50% (95% CI: 39 to 57) in Eastern Europe [101], whilst in Northern America and Australia, the HPV-attributable fractions (AFs) are 51% (95% CI: 41 to 57) and 41% (95% CI: 32 to 47%), respectively. The estimated proportion of OPCs attributed to HPV in Latin America is comparatively low at 13% (95% CI: 5 to 23), however, the prevalence of HPV has not been determined in many regions, including Brazil [102].

3.3.2.3. Laryngeal cancer

South America and Eastern Europe (e.g. Montenegro, Republic of Moldova and Hungary) are characterised by high incidence rates of laryngeal cancer (Figure 9 c.). Cuba has the highest incidence rates overall (8.9), and among males (16.2), whilst among females, the highest incidence rates correspond to Montenegro (3.4), Hungary (2.0) and Cuba (1.9). Tobacco use is the greatest single risk factor for the development of laryngeal cancer [103] [104], although alcohol consumption is also independently associated [104-106] and their combined use interacts in a multiplicative way [107], as will be discussed later in this chapter. The prevalence of tobacco smoking in Cuba is high at around 43% for men and 27% for women [108], which again likely accounts for the high incidence rates in this country. In fact, in 1995 and 2007, 82% and 84% of laryngeal cancers in Cuban males and 78% and 54% in Cuban females, respectively, were attributed to tobacco smoking [109]. As noted above, the prevalence of tobacco use in Eastern Europe is also high [110] (Figure 11). In Montenegro for example, where laryngeal cancer incidence is high in both males and females, 37.6% of men and 29.4% of women smoke.

3.3.2.4. Other HNC sites

Consistent with the high prevalence of tobacco and alcohol use, the country with the highest incidence of hypopharyngeal cancer, both overall and in sex-specific analyses, is Bangladesh (ASIRs: overall, 5.1; males, 8.6; females, 1.5). Hungary also has a high incidence among men (6.5).

A disproportionally high incidence of nasopharyngeal cancer (NPC) can be seen in Southeast Asia, especially China. For all of Southeast Asia combined, the incidence rate is 6.4 cases per 100,000 males and 2.0 per 100,000 females [111]. These rates may sound low but globally NPC is relatively rare with an estimated ASIR of 1.5 (2.2 for males and 0.8 for females) (Table 5). Intake of preserved foods such as salted fish and foods containing volatile nitrosamines at an early age has been linked to NPC risk in this population [112-114]. Other recognised risk factors include smoking, Epstein Barr virus, the use of traditional herbal medicines and occupational exposures to formaldehyde, wood dust, smoke, and chemicals [115]. High rates of NPC are also seen in Northern Africa. In Morocco for instance, NPC is the most common HNC, accounting for 7–12 % of all cancers in men [116]. Dietary factors such as the consumption of rancid butter, rancid sheep fat and preserved meats like quaddid (dried mutton stored in oil) have been found to be associated with an increased risk of NPC in this region [117].

Salivary gland cancer is a rare cancer globally (Table 6), however the highest incidence rates are found in Sweden (3.0) and Finland (2.0). The reasons for this are unclear as the aetiology of the disease is not well defined-largely because it is so uncommon. Unlike the majority of HNCs however, tobacco and alcohol use have not been strongly associated with salivary gland cancer [118 119]. The only well-established risk factor is Ionizing radiation, though some studies have suggested links between salivary gland cancer and occupational exposures, tobacco, ultraviolet light, and viruses [119-121]. The possibility of a viral aetiology seems plausible as salivary gland carcinoma is often observed in immunosuppressed individuals [119].

3.4. Trends in head and neck cancer incidence rates

Over the last few decades, there has been a dramatic change in HNC incidence trends by sub-site, country and gender and this has been attributed to a change in the prevalence of risk factors (i.e. tobacco use, alcohol consumption and HPV infection) [122 123].

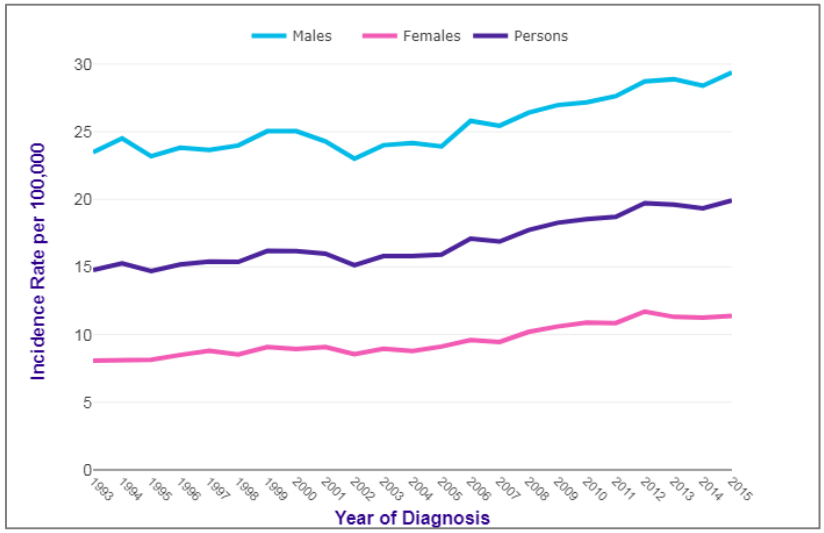
3.4.1. UK trends in incidence rates

Overall, HNC incidence rates have increased by a third (33%) in the UK since the early 1990s [64] (Figure 12). The increase has been larger in females than in males. Between 1993-1995 and 2015-2017 the ASI rates increased by 43% in females and 23% in males.

In contrast to the overall trend in HNC incidence rates, laryngeal cancer incidence rates have declined by 20% since the early 1990s, consistent with a reduction in the use of tobacco products though (Figure 13), rates have levelled off more recently [124] [125]. The overall increase in HNC incidence rates is largely attributed to a rise in oral cavity cancer rates. Between 2002 and 2011, there was a 25% increase in rates of oral cavity cancer, with rates having risen faster for women than for men [70]. One possible explanation for this is that, whilst smoking rates have decreased overall in the UK over the past 50 years or so, there was a lag in the 'adoption, diffusion and abatement' of cigarette use among women [126] [127] (Figure 13), meaning that these birth cohorts of women are only now reaching the high-risk age group for HNC. Specifically, consumption rates peaked among men in the mid-1960s but continued to increase among women up until the mid- to late-1970s [126] [127].

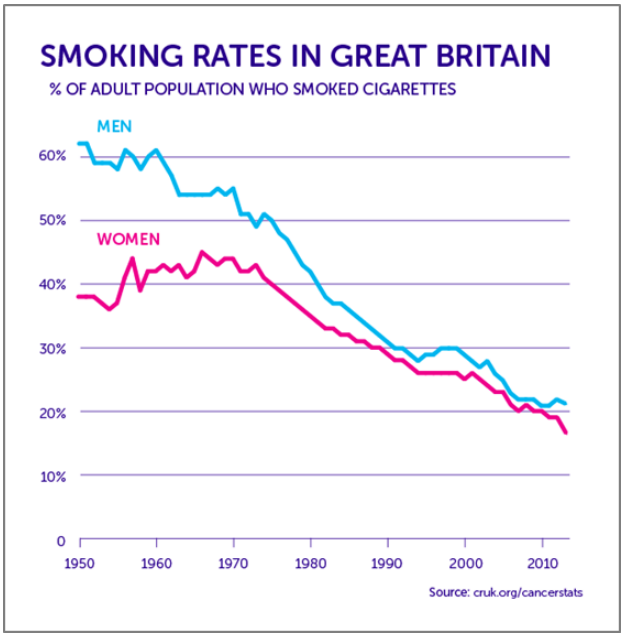
Another possible explanation for the recent increase in oral cavity cancer in the UK (and in several other European countries) is that alcohol consumption rates are currently at an historical high. Between the mid-1950s and late 1990s, the number of litres of alcohol consumed per person in the UK more than doubled [128], rising from around 4 litres per head in 1957 to 8.1 litres per head in 1997 [129]. Alcohol drinking is now on a downward trend, having peaked at 11.6 litres per head in 2005 (per capita consumption in the WHO European Region peaked at 12.2 litres in 2005 [130]), but rates remain high, especially among women [128]. For this reason, it has been suggested that the number of oral cavity cancers will continue to rise over the next decade or so [131].

Figure 12: Rise in HNC incidence in the UK, 1993-2017.



Credit: CRUK [132]

Figure 13: Trend in UK smoking rates among males and females, 1950-2010.



Credit: CRUK [132]*.*

3.4.2. Global trends in incidence rates

According to a review of 83 international cancer registries conducted by Simard *et al* (2014) [123], rates of oral cavity cancer increased among both men and women in 7 out of 37 countries between the periods of 1987-1993 and 1998-2002. These countries were located

in Europe and Asia and included the Czech Republic, Slovak Republic, Denmark, Estonia, Finland, the UK and Japan. Among men, the largest increase in rates was found in Finland (Rate ratio [RR]= 1.61, 95% CI: 1.39 to 1.86), while Spain had the largest increase in rates among women (RR= 2.23, 95% CI: 1.73 to 2.88). These findings are in-line with the literature which suggests a trend for increased rates of oral cavity cancer in some eastern and northern European countries [55 97 133 134], where similar patterns of cigarette smoking to the UK have been observed [135]. In other southern, central and eastern European countries, cigarette smoking is either still increasing among women or it has stabilised [136].

Several studies have also reported an increase in the incidence of OPC, predominantly in economically developed countries. In one study conducted by Chaturvedi *et al (2013)*, which used data from 70 registries in 23 countries across four continents, there were substantial increases in OPC incidence among men in several developed countries including the UK, USA, Canada, Australia, Denmark, Slovakia, and Japan; among women increases were observed exclusively in European countries (Denmark, Estonia, France, the Netherlands, Poland, Slovakia, Switzerland, and the UK) [137]. There were no significant increases in economically developing countries in South/Central America and Asia (e.g. Colombia, Costa Rica, Ecuador, India, the Philippines, Thailand). These findings are broadly consistent with those of Simard *et al*., who observed a trend for increased incidence of OPC in several eastern and northern European countries (Belarus, Czech Republic, Denmark, Finland, Norway, Sweden, and the UK), and a trend for decreased OPC incidence in China (Hong-Kong registry) and in parts of India (registry of Mumbai).

As mentioned previously, the main risk factors for OPC are tobacco, alcohol use and HPV infection. The recent increase in OPC incidence in developed countries has largely been attributed to an increase in the prevalence of HPV-associated cancers [123 137]. It is thought that generational changes in sexual behaviour, including the “sexual revolution” of 1960-1980, led to an increase in the transmission of oncogenic HPV [138], the details of which will be discussed shortly.

With respect to global trends in laryngeal cancer incidence, there is less evidence available. In the USA, the number of new laryngeal cancer cases has been falling by approximately 2.4% each year over the last 10 years. As is the case in the UK, this is consistent with a reduction in the prevalence of smoking [139]. According to The Surveillance, Epidemiology, and End Results (SEER) Program, the number of new laryngeal cancer cases fell from 5.2 per 100,000 American men and women in 1975 to 2.7 in 2015. In contrast, the above study by Simard *et al* found some evidence of an increase in laryngeal cancer instance in some

European countries (Belarus, Estonia and Latvia) over the last few decades. In this study however, the authors considered laryngeal cancer cases alongside other poorly-specified tumours of the lip/oral cavity, pharynx and hypopharynx (laryngeal cancers accounted for over two-thirds of all cases), and therefore additional studies focusing exclusively on laryngeal cancer instances are needed in order to confirm international trends.

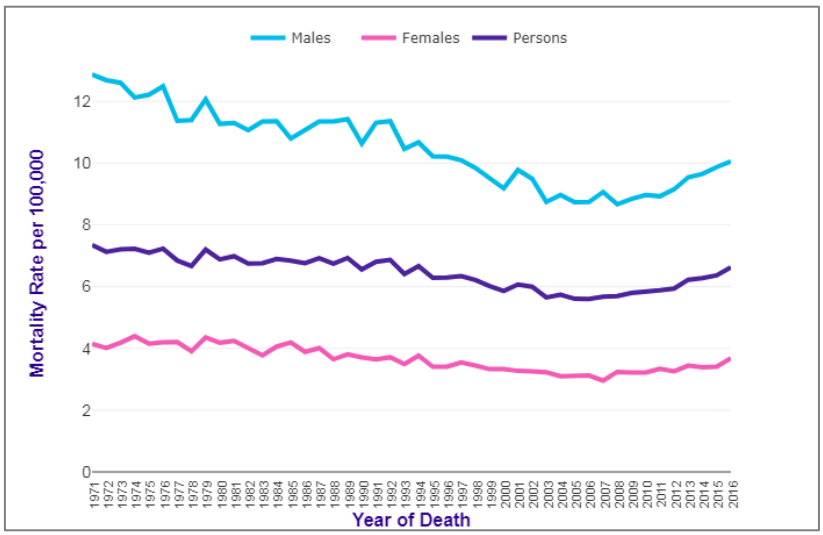
3.5. Mortality rates

3.5.1. UK mortality rates

According to CRUK figures, the number of people who died from HNC in the UK in 2016 was 4,046 [140]. Mortality rates in Scotland are higher than the UK average - 9.4/100,000 compared to 6.3/100,000 for England, Scotland and Wales combined. This reflects the higher incidence rate in this country. For the UK as a whole, between 1971-1973 and 2014-2016, there was an 11% decrease in European age-standardised mortality rates (Figure 14). However, mortality rates increased by 14% between the period of 2004-2006 and 2014-2016.

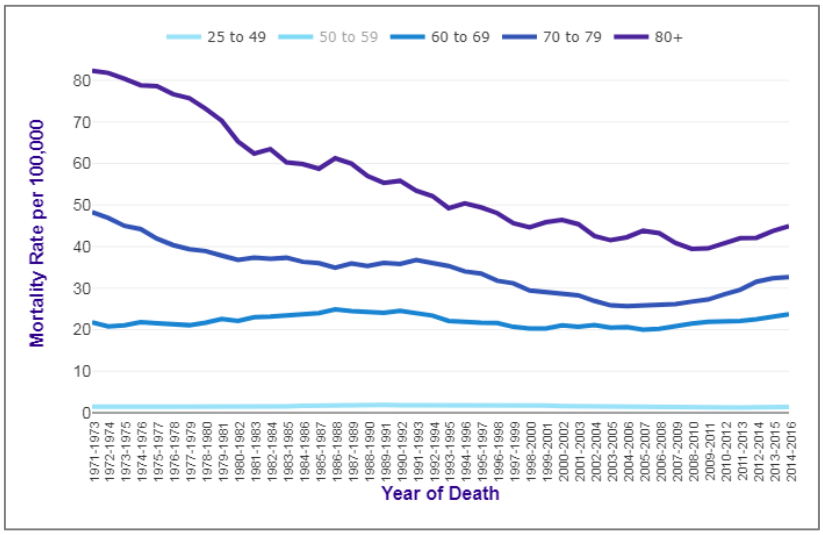
Mortality trends have varied by age, gender and sub-site. In UK males, rates have remained stable in the 25-49- and 60-69-year age groups, increased by 18% in the 50-59-year group an (Figure 15) [140]. In females, rates have decreased in most broad (adult) age groups, with the exception of the 60-69-year age group, in which rates have remained stable (Figure 16). Laryngeal cancer mortality rates fell by 33% between the period of 1990 to 2002 [125]. This may just reflect the fall in incidence rates in the UK, but it could also reflect changes in stage at presentation and treatment. By comparison, OPC mortality rates increased, though the increase in mortality rates is less significant than the increase in the incidence rates. It has been suggested that this is a result of the use of more effective treatments and combined therapy which have improved survival [125]. With regards to oral cavity cancer mortality, rates remained static during the same period.

Figure 14: Head and neck Cancer mortality trends over time in the UK (1971-2016).



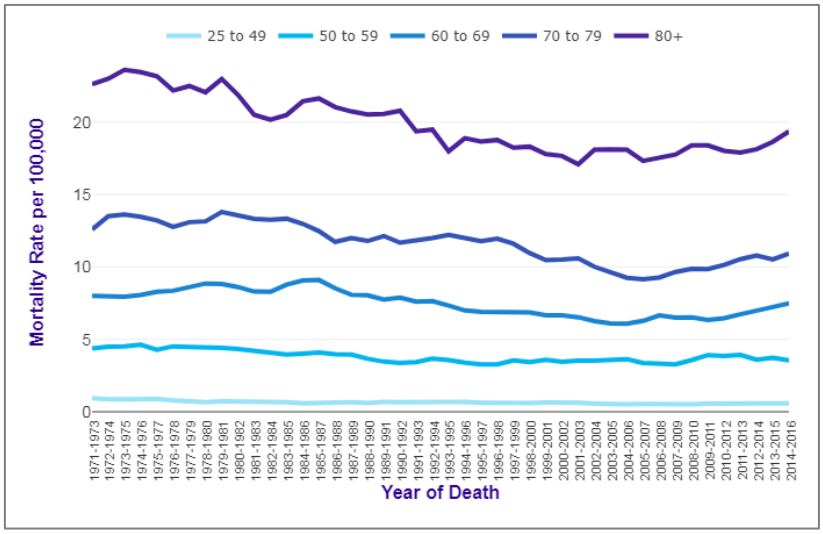
European age-standardised mortality rates per 100,000 population. Includes ICD C00-C14, C30-C32 Credit: CRUK [140]

Figure 15: Head and neck cancer European age-standardised mortality rates, by age, in UK males (1971-2016).



Includes ICD C00-C14, C30-C32. Credit: CRUK [140].

Figure 16: Head and neck cancer European age-standardised mortality rates, by age, in UK females (1971-2016).



Includes ICD C00-C14, C30-C32. Credit: CRUK [140].

3.5.2. Global mortality rates

Globally, it is estimated that HNC is responsible for around 450,000 deaths each year [62]. Of these deaths, approximately 177,000 are attributed to lip and oral cavity cancers (age-adjusted mortality rates of 2.0/100,000), 95,000 to laryngeal cancers (1.0), 73,000 to nasopharyngeal cancers (0.8), 51,000 to oropharyngeal cancers (0.6), 35,000 to hypopharyngeal cancers (0.4) and 22,000 to cancers of the salivary glands (0.2) [62]. It is important to note however that, in addition to the difficulties already highlighted at the beginning of this chapter regarding the availability of reliable incidence and mortality data, establishing cause of death in long-term survivors of HNC is problematic and therefore figures represent a best estimate of the number of deaths. Other competing causes of death in people with HNC include second primary malignancy, cardiovascular and pulmonary disease [141].

3.6. Major risk factors for HNC

The following section discusses the causal role of the major known risk factors for HNC. The WHO describe a risk factor as “any attribute, characteristic or exposure of an individual that increases the likelihood of developing a disease or injury” [142]. As mentioned already in this chapter, tobacco and alcohol intake are two of the most well-established risk factors for

HNC, most notably for cancers of the oral cavity, pharynx, and larynx, although cancers of the nasal cavity and paranasal sinuses have also been causally related to tobacco use [143]. Based on data from the INHANCE consortium of case-control studies (18 studies including 11,221 oral cavity, oropharyngeal and laryngeal cancer cases and 16,168 controls), Hashibe *et al* (2009) estimate that tobacco and alcohol use account for 33% and 4% of HNCs individually [144]. The effects associated with both tobacco and alcohol use were found to be greater than additive. That is, the combined effect of smoking and alcohol-drinking are more harmful than the individual effects of each exposure. Specifically, the total population attributable risk (PAR) for tobacco and alcohol combined was 72%. There were differences by subsite. The joint effects of tobacco and alcohol were responsible for 89% of laryngeal cancers, 72% of pharyngeal cancers and 64% of oral cavity cancers. Differences were also observed between genders. Among males, the total PAR for tobacco and alcohol was 74% compared to 57% among women.

Although tobacco and alcohol use are responsible for the majority of HNCs, a substantial proportion of cases cannot be attributed to these exposures. Tumours may occur in non-smoking, non-drinking individuals [145-147] and only a fraction of those who do smoke and drink develop tumours. This suggests that additional risk factors may be independently involved or act as co-factors for the development of HNC. HPV has already been mentioned as an important aetiological factor for OPC; its role in development of HNCs outside the oropharynx is less well-established [148]. Other recognised risk factors that will be discussed further include Epstein-Barr virus (EBV), low fruit and vegetable intake, low physical activity levels, poor oral and dental hygiene, a family history of cancer and genetic factors. Additional risk factors, including prolonged sun exposure [149] and certain occupational exposures [150-152] do exist, but will not be considered further in this thesis.

3.6.1. Tobacco use

There are a wide variety of tobacco products, including rolls of tobacco that are smoked (e.g. cigarettes, cigars, bidi), oral preparations which are chewed, held in the mouth or placed in the nose (e.g. snuff, snus, betel quid), and pipes, including water pipes [153]. Within the UK, cigarette smoking, hereafter referred to as "smoking," remains the most widely used form of tobacco use. In a pooled analysis of 15 case-control studies that included 10,244 HNC cases and 15,227 controls, Hashibe and colleagues found that, among never drinkers, cigarette smoking was associated with a two-fold increased risk of HNC (OR for ever versus never smoking= 2.13, 95% CI: 1.52 to 2.98) [154]. There was further evidence of dose-response relationships for frequency, duration, and number of pack-years of smoking. Even

for the lowest category of smoking (1-10 cigarettes per day) the risk was almost two-fold (OR = 1.82; 95% CI:1.28 to 2.59). When the risks were examined for specific subsites, laryngeal cancer was most strongly associated with cigarette smoking (OR= 6.84; 95% CI: 4.25 to 11.01) followed by OPC (OR= 2.02; 95% CI: 1.34 to 3.05) and oral cavity cancer (OR= 1.35; 95% CI: 0.90 to 2.01). In a separate analysis of data from the INHANCE consortium (1761 laryngeal, 2453 pharyngeal and 1990 oral cavity cancer cases and over 8,000 controls), Lubin *et al* (2009) assessed the risk by total exposure (measured in pack-years of smoking) and its modification by exposure rate (number of cigarettes per day). Their results suggest that smoking more cigarettes per day for a shorter period of time is less harmful than smoking fewer cigarettes per day for a longer period of time (above 15 cigarettes/day). Quitting smoking has been reported to reduce the risk of HNC [155-161]. Using data from 17 studies on smoking cessation (12 040 cases and 16 884 controls), Marron *et al* (2010) found that, compared to current smokers, those who abstained from smoking for 1-4 years experienced a reduction in risk of HNC (OR= 0.70; 95% CI: 0.61 to 0.81). For people who quit smoking for 20 years or more, the reduced risk of HNC was similar to that of never-smokers (OR= 0.23, 95% CI: 0.18 to 0.31) [156].

The smoke exhaled by a smoker, termed the mainstream smoke, contains around 4,800 identified compounds, 69 of which are known to be carcinogenic (cancer-causing) [162]. Some of these carcinogens occur naturally in the tobacco plant, whilst others are formed during combustion of the cigarette. A detailed discussion of the mechanisms by which these compounds induce carcinoma is beyond the scope of this review. However, the formation of DNA adducts are recognised as being central to the process of carcinogenesis [163 164]. When tobacco carcinogens (or their metabolites) bind covalently to DNA to form DNA adducts, this disrupts the double helical structure of DNA and interferes with its replication. Cells have evolved a number of cellular repair systems, including base and nucleotide excision repair, mismatch repair, and double-strand break repair, which can remove DNA adducts to conserve the normal DNA structure [165]. If repair enzymes are unable to function efficiently however, and the DNA adducts persist, this can result in miscoding (e.g. insertion of the incorrect base) and permanent mutations. Should these mutations occur at susceptible sites in the genome, such as in proto-oncogene (e.g. *K-ras*) or tumour suppressor genes (e.g. p53), this can lead to loss of normal growth control and ultimately, the development of tumours [163 166]. The situation is made worse by the presence of other chemicals in cigarettes, such as chromium, arsenic and nickel. Chromium allows compounds like benzo(a)pyrene to bind to DNA more strongly, whilst arsenic and nickel interfere with DNA repair pathways [167].

3.6.2. Alcohol use

Among never users of tobacco, Hashibe *et al* (2007) found that alcohol drinking was associated with a two-fold increased risk of HNC, but only in people who consumed three or more alcoholic beverages per day (OR= 2.04; 95% CI: 1.29 to 3.21 vs never drinkers) [154]. When the analysis was stratified by cancer subsite, there were monotonic relationships between pharyngeal (including oropharynx and hypopharynx) and laryngeal cancer risk and frequency of alcohol consumption. For pharyngeal cancer, increased risks were observed for those drinking one to two drinks per day (OR= 1.66; 95% CI: 1.18 to 2.34), while for laryngeal cancer, an increased risk was observed for those drinking five or more drinks per day (OR= 2.98; 95% CI: 1.72 to 5.17).

In contrast to their results for smoking, Lubin *et al* observed that, for subjects consuming 10 drinks per day or less (above 10 drinks per day, data were limited), the strength of the HNC risk association with total exposure (years of alcohol drinking) increased with increasing exposure rate (number of alcoholic drinks per day), suggesting that greater drinks per day for a shorter period of time was more harmful than fewer drinks per day for a longer period of time [168]. Excess odds ratio (EOR) per drink year estimates varied by site. The risk was greatest for oral and pharyngeal cancer.

Marron *et al* (2010) found that quitting alcohol drinking was also associated with a reduction in the risk of HNC (OR= 0.60; 95% CI: 0.40 to 0.89 compared with current drinking), but the benefits were only observed after ≥20 years of abstaining (compared to 1 to 4 years for smoking cessation) [156]. The risk reversal after quitting drinking for ≥20 years is observed across all subsites of HNC (oral cavity, oropharynx and larynx).

Using data from the INHANCE consortium, Purdue *et al* (2009) sought to understand the risks associated with consuming different types of alcoholic beverages [169]. Specifically, they calculated associations of HNC with different measures of beverage consumption in people who drank beer only (858 cases, 986 controls), spirits only (499 cases, 527 controls), and wine only (1,021 cases, 2,460 controls), using never-drinkers (1,124 cases, 3,487 controls) as the reference category. The authors observed similar associations with HNC among beer-only and spirit-only drinkers; among beer-only drinkers, ORs were 1.6, 1.9, 2.2, and 5.4 for subjects who drank ≤5, 6–15, 16–30, and >30 ethanol-standardized drinks per week ($p_{trend} < 0.0001$); among spirit-only drinkers, the corresponding ORs were 1.6, 1.5, 2.3, and 3.6 ($p < 0.000$). For wine-only drinkers, ORs for moderate levels of consumption (≤5 and 6–15 drinks per week) were close to null and increases in risk were only observed at higher

consumption levels (ORs: 1.9 and 6.3 for 16–30 and >30 standard drinks per week [$p < 0.000$], respectively). The authors provide several possible explanations for the observed weaker association with moderate wine consumption. Firstly, it may be the result of residual confounding e.g. wine consumption has been associated with higher intake of a healthy diet but they were unable to adjust for this in their analysis [170-172]; secondly, they propose a possible "alcohol washing effect" (i.e. the ingestion of foods could modify the absorption and effect of alcohol on oral mucosa [173]) since wine is more frequently consumed with food than other alcoholic beverages; and finally, it is possible that the carcinogenic effects of alcohol are offset by other anticarcinogenic compounds found in wine e.g. resveratrol [174].

The main constituents of alcohol-containing beverages are ethanol (C_2H_5OH), water and glucose. Whilst alcohol itself is not a carcinogen, its use can increase cancer risk in two main ways. Firstly, alcohol acts as a solvent to increase the permeability of cellular membranes, thereby allowing other carcinogenic compounds (namely those found in tobacco smoke), to penetrate the mucosal surfaces of the upper digestive tract [175]. Secondly, Acetaldehyde, the first metabolite of ethanol, has wide-ranging genotoxic and carcinogenic effects. Amongst other things, acetaldehyde has been shown to: interfere with DNA repair mechanisms, for example by inhibiting the enzyme O6-methylguanine transferase, which is responsible for repairing DNA damage caused by alkylating agents; induce sister chromatid exchanges and gross chromosomal aberrations; introduce point mutations, for instance in human lymphocyte cells; and form DNA adducts [176-178]. In addition to this, chronic alcohol consumption is associated with immune suppression, which may help facilitate tumour spread [175 177].

3.6.3. Human papilloma virus infection

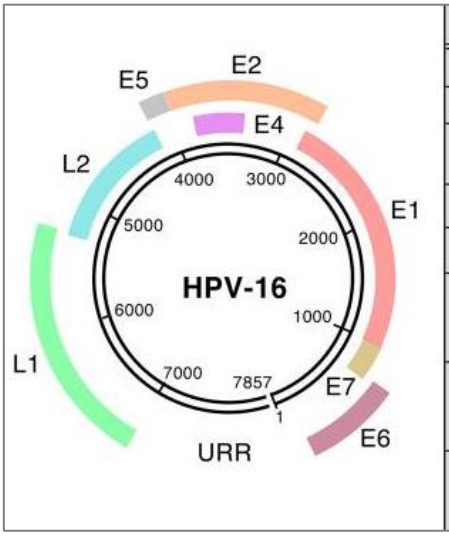
HPV is one of the most common sexually transmitted diseases. Infections are usually asymptomatic, and the virus is cleared from the body within a couple of years. However, for some people, genital HPV infection can result in clinical disease, including anogenital warts, cervical cancer and other anogenital cancers (e.g. cancers of the penis, vagina and anus) [179]. In 2007, in the ninetieth volume of *the IARC Monographs on the Evaluation of Carcinogenic Risks to Humans,* the International Agency for Research on Cancer (IARC) concluded that there was sufficient epidemiological evidence to show that HPV is a causative factor for malignant transformation in a subset of HNCs [179]. They found that HPV DNA was consistently detected in a large proportion of OPCs, with an estimated average prevalence of 35%. By comparison, HPV DNA was detected in some oral cavity cancers, but

the range of detection was wide (average of approximately 25%) and there was limited evidence to support the carcinogenicity of HPV in the larynx. The working group point out that it is not always possible to determine the precise anatomical location of the primary tumour and misclassifications could result in distorted prevalence estimates for individual sub-sites. For example, the base of the tongue is inconsistently grouped as either oral cavity or oropharynx. Besides differences in anatomical grouping, sample numbers, differences in sampling techniques (i.e. frozen, formalin-fixed or paraffin-embedded sections, scraping or oral rinses), and different detection methods could all influence estimated prevalence rates (133). The various methods of HPV detection will be discussed later in this chapter.

More than 150 different sub-types of HPV are known to exist, but the vast majority of these are “low- risk”, meaning that they are of low carcinogenic potential [180] [181]. There are 15 “high-risk” HPV sub-types and of these, HPV 16 is the most common type associated with HNC. Around 90% of HPV-associated HNCs are attributed to HPV 16 [182-184], with HPV18, 31, 33 and 35 making up the remaining cases [182]. The majority of these HPV-associated HNCs develop in the lymphoid tissue of the oropharynx, which includes the lingual and palatine tonsils [185].

All HPV subtypes share a common genetic structure [179] (Figure 17). They have a small, circular, double stranded DNA genome of approximately 8000 base pairs in length, bound within a protein capsid [179]. The genomes comprise eight open reading frames (ORFs), each of which is transcribed from a single DNA strand. The ORFs correspond to three functional regions; an “early” region that encodes regulatory proteins E1-E7, a “late” region that encodes capsid proteins L1-L2, and a largely non-coding region that is referred to as the long control region (LCR) [179]. The LRC contains *cis* elements that are necessary for viral DNA replication and transcription. The oncogenic nature of HPV is attributed to the presence of high-risk E6 and E7 proteins, which cooperate to suppress normal cell cycle controls. Further details of the role of HPV in the pathogenesis of HNC are provided later in this chapter in the section *HPV detection methods in HNC*.

Figure 17: HPV-16 genome



Gene	Function
L1	Major capsid protein.
L2	Minor capsid protein.
E1	Transcription factor, helicase activity. Mediates episomal DNA replication.
E2	Transcription factor. Regulates viral copy number.
E4	Facilitates virion release.
E5	Stimulates cell proliferation and prevents differentiation. Downregulates surface MHC class I expression.
E6	Deregulates cell cycle control through p53 inactivation/degradation. Induces malignant transformation together with E7.
E7	Keeps cells active in the cell cycle through Rb inactivation. Induces malignant transformation alone and together with E6.

Image source: Riemer et al (2010) [186].

As is the case with cervical and urogenital cancers, HPV infection of the oropharynx (and oral cavity), has been shown to be associated with high-risk sexual behaviours, particularly oral sex [185]. In a hospital-based, case-control study of 100 people with newly diagnosed OPCs and 200 controls, D'Souza *et al* (2007) found that the risk of developing OPC was three times greater for people with high lifetime numbers of vaginal- and oral sex partners (HR= 3.1; 95% CI: 1.5 to 6.6 for people who had had 26+ partners versus people who had has 0-5 partners and HR= 3.4; 95% CI: 1.3 to 8.8 for people who had had 6+ oral sex partners versus none) [185]. For people with HPV-16 positive tumours, the risk increased over four-fold and eight-fold, respectively (HR= 4.2; 95% CI: 1.8 to 9.4 for people who had had 26+ vaginal-sex partners and HR= 8.6; 95% CI: 2.2 to 34.0 for people who had had 6+ oral-sex partners:). In the same study, the authors sought to examine whether combined exposure to HPV and tobacco and/or alcohol further increased OPC risk. They found no evidence of a synergistic effect; however, some other studies suggest that smoking has the potential to promote infection and viral persistence, possibly increasing the carcinogenetic effect of HPV [187]. In one US study, which included 284 OPC cases and 477 controls, the joint effect of smoking and HPV-16 seropositivity (measured using Enzyme Linked Immunosorbent Assay [ELISA] for detection of antibodies to L1) was associated with an 8.5-fold increased risk of OPC (95% CI: 5.1 to 14.4), which was much stronger than the predicted sum of individual associations with current smoking (OR= 3.2; 95% CI: 2.0 to 5.2) and HPV-16 seropositivity (HR= 1.7; 95% CI: 1.1 to 2.6) [188].

3.6.4. Epstein-Barr virus

EBV, the same virus that causes glandular fever, has been etiologically linked to NPC, especially the non-keratinizing form of NPC [189] [190]. Infection with EBV alone is not sufficient to cause cancer. Indeed, sero-epidemiological studies indicate that globally more 90% of the adult population are infected [191] [192] and yet only a small fraction of these infections will ever progress to cancer. The reasons why some latent infections progress to cancer are not yet clear, but genetic susceptibility, environmental factors and interactions with other pathogens may all play a role.

3.6.5. Diet and nutrition

Epidemiological studies suggest a link between diet and HNC [193-198]. In particular, a high intake of fruits and vegetables has been associated with a reduction in risk. In a study of almost 350,000 subjects included in the Prospective European Investigation into Cancer and Nutrition (EPIC), for example, an 80g/day increase in fruit and vegetable intake was associated with a 10% decrease in the incidence rate of upper aerodigestive tract (UADT) cancers (including pharynx, larynx, and oesophagus; RR= 0.91; 95% CI:0.83 to 1.00) [197]. Similarly, Edefonti *et al* (2012) found evidence of an inverse association between an "antioxidant vitamins and fibre" dietary pattern, which is rich in fruit and vegetables, and risk of oral and pharyngeal cancer (OR: 0.57; 95% CI: 0.43 to 0.76 for the highest versus the lowest quintile) in a study that included 2,452 cases and 5,013 controls [193].

There is evidence to suggest that fruit and vegetables may protect against the carcinogenic effects of tobacco and alcohol [195] [196]. In a pooled analysis of case-control studies conducted in 9 countries worldwide (1,670 cases and 1,732 controls), Kreimer *et al* (2006) found that high (versus low) fruit and vegetable intake significantly reduced oral and pharyngeal cancer risk, but only among ever-tobacco users (OR = 0.4; 95% CI: 0.3 to 0.6 for ever-tobacco users and 1.1; 96% CI: 0.6 to 2.0 for never-tobacco users) and ever-drinkers (OR = 0.4; 95% CI: 0.3 to 0.6 for ever-drinkers and 1.0; 95% CI: 0.6 to 1.6 for never-drinkers). The protective effects of fruits and vegetables may be attributed in part to carotenoids, which are recognized antioxidants with anti-mutagenic and immuno-regulatory activities [195].

3.6.6. Body mass index

Most previous studies report that people with a low body mass index (BMI) i.e <18.5 experience a higher risk of HNC compared to those of normal body mass i.e. 18.5 to <25.0 [196] [199-205]. However, as Gaudet *et al* (2010) [205] point out, it is difficult to assess the

independent effects of BMI on HNC risk as BMI is so strongly associated with smoking and alcohol drinking [206-208], such that even after adjusting for tobacco and alcohol consumption, the influence of residual confounding cannot easily be dismissed. Lubin *et al* (2010) examined the relationships between BMI, smoking, and alcohol consumption in the INHANCE consortium and found that the relationships differed by anatomical sub-site. Low BMI was found to significantly enhance smoking and alcohol-related associations for oral cavity and pharyngeal cancer but not for laryngeal cancer [209]. For oral cavity/pharyngeal cancer, BMI modified the strength of the relation between cancer risk and drink-years, specifically whilst the impact of BMI on the relation between disease risk and drink-days were similar across all categories of BMI.

3.6.7. Physical activity

There is increasing evidence to suggest that physical activity (PA) may lower cancer risk but so far, the data for HNC is limited and conflicting. In one study conducted by Nicolotti *et al* (2011), which included 2,289 HNC cases and 5,580 controls, subjects who participated in moderate recreational PA (definitions varied by study) experienced a 22% reduced risk of HNC in multivariable-adjusted models (OR= 0.78; 95% CI: 0.66 to 0.91), compared to those in the none/low PA group [210]. In another prospective study (NIH-AARP Diet and Health Study) that followed 487,732 subjects (1,249 HNC cases) for eight years, the relation between PA and HNC risk substantially attenuated and became statistically non-significant after adjustment for smoking (additional adjustment for other potential confounding variables had little influence on the risk estimate) [211]. These findings led the authors to conclude that PA is unlikely to play an important role in the development of HNC.

3.6.8. Oral hygiene

People with poor oral health, i.e. individuals with several missing teeth, bleeding gums or chronic infection, have been found to have a higher risk of developing HNC, though the effects are often modest [212-216]. The processes underlying the apparent associations between oral and dental hygiene and HNC risk have yet to be fully elucidated, but plausible mechanisms involve: systemic inflammation and the release of cytokines and other inflammatory mediators [214]; bacterial overgrowth, which may contribute to the formation of endogenous nitrosamines [214 217 218] and increased formation of the highly carcinogenic intermediate acetaldehyde from ethanol [214 216 219]; and the facilitation of oral HPV infection, for example, by providing an entry point for the virus [220].

3.6.9. Socioeconomic position

A 2015 study, which used pooled data from 31 studies (n=23,964 HNC cases) in 27 countries, examined the effects of income and education on HNC occurrence [221]. The findings were that fewer years of education and lower income were associated with an increased risk of disease development. Overall, low education, as defined by an International Standard Classification of Education (ISCED) score of 0–1 (no education, or completed the first stage of basic education, or at most primary education) was associated with a 2.5-fold increased risk of HNC (OR = 2.50; 95% CI = 2.02 - 3.09, p<0.001) relative to high educational attainment (ISCED 5-6) in models that adjusted for age, sex and centre (for multiple-centred studies). Those in the intermediate level of educational attainment had an intermediate increased risk (OR = 1.80; 95% CI = 1.57–2.07, p<0.001). Low relative to high household income (based on strata) was associated with a similar increased risk of head and neck cancer (OR = 2.44; 95% CI = 1.62–3.67).

The authors estimated the unexplained or "direct" effect of low socioeconomic status (SES) on HNC risk. They found that the proportion of the increased risk estimate associated with low educational attainment not explained by smoking alone was 58%; by smoking and alcohol combined was 31%; by smoking, alcohol and diet was 29% and by smoking, alcohol, diet and other tobacco use was 23%. For income, 39% of this risk was not explained when adjusting for smoking and alcohol (income models were not adjusted for diet and other tobacco use in the main analysis). When the analysis was restricted to those who never smoked and never drank alcohol, lower educational attainment remained associated with a >50% increased risk (OR = 1.61; 95% CI = 1.13–2.31). Taken together, these results suggest that risk associated with low SES cannot entirely be explained by differences in the distributions of behavioural risk factors.

3.6.10. Family history of head and neck cancer

The results of three record linkage studies [222-224] suggest that the standardised incidence ratio (SIR), which is the ratio of the observed number of cases to the expected number of cases [225], for developing HNC among people with a family history ranges between 1.4 and 8.0. Again, using data from the INHANCE consortium (12 case-control studies including 8,967 HNC cases and 13,627 controls), Negri *et al* (2009) established that having a first-degree relative with HNC was associated with an almost two-fold increased risk (OR= 1.7; 95% CI: 1.2 to 2.3). The risk was highest when the affected relative was a sibling (OR sibling: 2.2; 95% CI: 1.6 to 3.1; OR parent: 1.5; 95% CI: 1.1 to 1.8). In subjects with a family

history who were also alcohol and tobacco users, the risk increased seven-fold (OR= 7.2; 95% CI: 5.5 to 9.5).

3.6.11. Genetic risk factors

Although most HNCs are caused by carcinogens or viral infection, a small percentage of cases are familial in nature. Inherited conditions which are known to increase HNC risk include Fanconi anaemia, ataxia telangiectasia, Bloom's syndrome and Li–Fraumeni syndrome [226].

It is widely hypothesised that variations in genes involved in pathways such as carcinogen metabolism DNA repair and cell cycle control may increase the risk of tobacco and alcohol associated carcinogenesis [227], although the results of candidate gene studies have been mixed. For further discussion, the reader is directed towards reviews by Hiyama *et al* (2008) and Cadoni *et al* (2012) [228] [229]. Among the most extensively studied genes are those in the alcohol dehydrogenase (ADH) gene family. Polymorphisms (i.e. genetic variants) in *ADHB1* in particular have been strongly associated with HNC susceptibility. In an analysis of five ADH variants, which included 3,876 upper aerodigestive tract (UADT) cancer cases (790 cancers of the oral cavity or pharynx, 1,659 cancers of the hypopharynx or larynx and 427 cancers of the oesophagus) and 5,278 controls, the A-allele of rs1229984 (also known in the literature as *ADH1B*2*), which confers fast metabolism of ethanol to acetaldehyde, was shown to be protective against aerodigestive cancer [230]. Specifically, the reduction in risk was between two and three-fold for oral and pharyngeal cancer (OR= 0.45; 95% CI: 0.35 to 0.57) and 30% for laryngeal cancer (OR= 0.71; 95% CI: 0.57 to 0.88). The effect became more apparent with increasing alcohol consumption (*P for trend* = 0.0002) [230]. A summary of some of the other most commonly cited polymorphisms is provided in Table 7.

Subsequent genome-wide association studies (GWAS) have validated the single nucleotide polymorphism (SNPs) in *ADH1B* (reference SNP cluster ID: rs1229984) [231-233] and have identified a number of additional susceptibility loci to date, including other variants in the *ADH gene cluster and in* Aldehyde dehydrogenase 2 (*ALDH2)*. An overview of GWAS findings is provided in Table 8. In one study, McKay *et al*. identified five common variants associated with UADT cancer susceptibility in the INHANCE consortium (replication n= 6,514 UADT cancer cases and 7,892 controls of European descent), including three located within the alcohol dehydrogenase genes (*ADH1B*, *ADH1C*, *ADH7*) that have been described previously [230] and two novel variants, a 4q21 variant (rs1494961), which resides near DNA repair related genes *HEL308* and *FAM175A,* and a 12q24 variant (rs4767364) located in a

region that contains multiple genes including the (*ALDH2*) gene [231]. The rs1229984 (4q23, *ADH1B*) association was subsequently validated by Leusseur *et al*. in a GWAS meta-analysis of overall oral and pharyngeal cancer (oral, oropharynx, hypopharynx and overlapping cancers; n = 6,034), and in site-specific oral cancer (n = 2,990) and OPC (n = 2,641) analyses [232]. In addition, this study identified a large association signal at 6p21.32, within the human leukocyte antigen (HLA) region in overall and oropharyngeal-specific meta-analyses.

Table 7: Some of the most intensely studies genetic polymorphisms linked to HNC risk.

Gene	Activity	Association with HNC	Ref
Carcinogen metabolising genes			
Glutathione S-transferase (*GSTM1*)	GSTs are phase II xenobiotic metabolizing enzymes. They are involved in catalysing the conjugation reactions of reactive intermediates of electrophilic compounds with cytosolic glutathione (GSH) [234].	Null *GSTM1* genotype (which lacks the corresponding enzyme function) is associated with increased risk vs. the positive genotype.	[235]
Glutathione S-transferase (*GSTT1*)		Null *GSTT1* genotype (which lack the corresponding enzyme function) is associated with increased risk vs. the positive genotype.	[235]
Glutathione S-transferase (*GSTP1*)		Ile/Val and/or Val/Val genotypes (*GSTP1* codon 105) are associated with increased risks vs. the Ile/Ile genotype.	[235]
Cytochrome P450 2E1 (*CYP2E1*)	*CYP2E1* is one of the main constituents of the microsomal ethanol oxidation system, which breaks down ethanol to acetaldehyde and generates reactive oxygen intermediaries. *CYP1A1* encodes enzymes that are active in the metabolism of several tobacco-derived pro-carcinogens, including polyaromatic hydrocarbons, N-nitrosamines and aromatic amines.	c2/c2 genotype associated with increased risk vs. c1/c1 genotype.	[236]
Cytochrome P450 1A1 (*CYP1A1*)		Ile/Val and/or Val/Val genotypes (which increase enzymic activity) are associated with increased risks vs. the Ile/Ile genotype. MspI polymorphism associated with increased risk of HNC.	[235] [237]
N-acetyltransferase 2 (*NAT2*)	A key enzyme involved in the metabolism of numerous aromatic and heterocyclic amine carcinogens [228]. Polymorphic changes regulating the acetylation capacity of *NAT2* have been implicated in the risk of developing several tobacco-associated cancers [238-240].	Slow acetylation genotype associated with increased risk vs. rapid genotype.	[241-247]
Human microsomal epoxide hydrolase (*EPHX1*)	The human microsomal epoxide hydrolase (mEH) enzyme, which is encoded by *EPHX1,* is involved in the metabolism of many potentially carcinogenic or genotoxic	Tyr/His genotype (which corresponds with low enzyme activity) is associated with lower risk vs. the Tyr/Tyr genotype.	[249-252]

	epoxidesincluding those derived from the oxidation of polyaromatic hydrocarbons [228] [248].		
Alcohol metabolising genes			
Aldehyde dehydrogenase 2 (*ALDH2*)	In the first stage of alcohol metabolism, ADH metabolises ethanol to acetaldehyde, a known carcinogen [253]; In the second stage, ALDH rapidly breaks down acetaldehyde to another less toxic by-product called acetate. Polymorphisms of both ADH and ALDH genes are known to determine blood acetaldehyde concentrations after drinking [254] [255].	Carriers of the "slow metabolising" allele *ALDH2*2* are at increased risk.	[255]
Alcohol dehydrogenase isoenzymes (*ADH*)		Carriers of the "fast metabolising" allele *ADH1B*2* (rs1229984) are at reduced risk compared to individuals with the low-activity *ADH1B*1/*1* (wild-type) genotype.	[230] [255]
DNA repair genes			
Xeroderma pigmentosum complementary group D (*XPD*)	One of the major genes involved in nucleotide excision repair [228], which is the main pathway responsible for removing lesions or sites of damage in the DNA, [256].	Individuals homozygote for the variant genotype 22541AA (XPD codon 156) have decreased risk compared to 22541CC homozygotes. Individuals homozygous for the variant genotype A35931CC (XPD codon 751) have increased risk compared to A35931AA homozygotes.	[257-260] [257-262]
Cell cycle repair genes			
P53	Acts as the major cellular "gatekeeper" for growth and division.	Conflicting evidence to suggest that a functional polymorphism which codes either arginine (Arg) or proline (Pro) at codon 72 of exon 4 is associated with HNC risk. Most studies suggest that the Pro/Pro genotype is associated with increased risk vs. the Arg/Arg genotype.	[228]
Cyclin D1 (*CCND1*)	Cyclin D1 (*CCND1)* is an important regulator of the G1 phase of the cell cycle. It is amplified in 30% to 50% of HNCs [263] [264].	A G870A polymorphism, which creates an alternative splice variant of the *CCND1* gene, is associated with increased risk.	[265-270]

Table 8: HNC risk loci identified in GWAS.

Cancer	Region	Reference SNP	Gene	Minor allele	p-value*	Ref
OC, OPC	4q23	rs1229984	*ADH1B*	T	$2.29x10^{-15}$	232
OC, OPC	6p21.32	rs3828805	*HLA-DQB1*	T	$3.35x10^{-13}$	232
OC, OPC	10q26.13	rs201982221	*LHPP*	A	$1.58x10^{-9}$	232
OC, OPC	11p15.4	rs1453414	*OR52N2/TRIM5*	C	$4.78x10^{-8}$	232
OC	2p23.3	rs6547741	*GPN1*	G	$3.97x10^{-8}$	232
OC	4q23	rs1229984	*ADH1B*	T	$1.09x10^{-9}$	232
OC	5p15.33	rs10462706	*CLPTM1L*	T	$5.54x10^{-10}$	232
OC	9p21.3	rs8181047	*CDKN2B-AS1*	A	$3.80x10^{-9}$	232
OC	9q34.12	rs928674	*LAMC3*	G	$2.09x10^{-8}$	232
OPC	4q23	rs1229984	*ADH1B*	T	$8.53x10^{-9}$	232
OPC	6p21.32	rs3828805	*HLA-DQB1*	T	$2.21x10^{-12}$	232
UADT	4q21	rs1494961	*HEL308*	C	1×10^{-8}	231
UADT	12q24	rs4767364	*ALDH2*	A	2×10^{-8}	231
UADT	4q23	rs1229984	*ADH1B*	T	1×10^{-20}	231
UADT	4q23	rs1573496	*ADH7*	C	9×10^{-17}	231
UADT	4q23	rs698	*ADH1C*	C	3×10^{-7}	231
Lar	6p21.33	rs2857595	*AIF1*	A	2.43×10^{-15}	233
Lar	11q12.2	rs174549	*FADS1*	A	1.00×10^{-20}	233
Lar	12q24.21	rs10492336	*TBX5*	A	4.48×10^{-14}	233

** P-values presented relate to discovery population. Abreviations:* ***SNP****, single nucleotide polymorphism.*

3.7. Multi-omics profiling in HNC

HNC (and cancer in general) is a complex disease involving alterations at multiple molecular levels including the genome, epigenome, transcriptome, proteome and metabolome. As such, there has been a growing trend towards integrating multiple "omics" technologies into HNC cancer research [271] [272] in order to try to obtain a better understanding of the systemic processes that drive cancer initiation and maintain tumorigenesis. This thesis will be considering epigenetic and metabolomic biomarkers specifically but a brief description of the genetic changes that characterise HNC will also be presented below, in order to provide a more complete picture of the molecular profile of the disease. Details of the different epigenetic and metabolomic platforms themselves will be provided in the next chapter.

3.7.1. Genetic landscape

Exome-sequencing studies (i.e. studies that sequence the protein-coding regions of the genome [273]), have revealed that HNC has a relatively high mutational load, ranking eighth highest in an analysis of 27 cancer types [274]. The mutational rate (i.e. total number of mutations per coding area of a tumour genome) is broadly comparable to that of other smoking-related malignancies such as lung and oesophageal cancer [275]. One study reported significantly lower mutation rates in HPV-associated HNCs, hereafter referred to as HPV-positive cancers; in fact, the frequency of mutation was around half that of non-HPV associated tumours, hereafter referred to as HPV-negative cancers (mean mutation rate of 2.28 mutations per megabase [Mb] compared with 4.83 mutations/Mb; $p = 0.004$ for [275] for HPV-negative tumours). The authors suggest that higher genome instability in HPV-negative cancers could lead to a higher risk of developing treatment resistance and this may go part way to explaining why people with HPV-positive tumours have improved survival compared to people with HPV-negative cancers [276].

In the largest analysis of its kind to date (n=279), the Cancer Genome Atlas (TCGA) identified eleven genes that were consistently mutated in HNC [277]. The mutational frequency of several of these genes, including *TP53, PIK3CA, NOTCH1, TP63* and *CDKN2A,* were comparable in a later analysis by Perdomo *et al* (2016), which included 180 paired samples diagnosed with HNC in two high incidence regions of Europe and South America [277]. Among inactivating mutations, four genes identified in the TCGA study were found to segregate exclusively or predominantly in non HPV-associated tumours; these included *TP53* and *CDKN2A*, which are associated with cell cycle control, and *FAT1* and *AJUBA*, which are

linked to Wnt/β-catenin signalling. *TP53* mutations have been associated with poor clinical outcomes, including inferior therapeutic response [276]. Table 9 provides an overview of the mutational frequencies and types of alterations found in HNC, stratified by HPV status. For a more in depth discussion of HNC genomics see Beck *et al* (2016) [278]

3.7.2. Epigenetic signatures

Accumulating evidence suggests that epigenetic mechanisms, i.e. mechanisms that produce changes in gene expression without altering the underlying genetic code, play an important role in carcinogenesis, tumour progression, and resistance to therapy [279 280]. DNA methylation (DNAm), which involves the addition of a methyl group to the fifth carbon of the cytosine DNA base, producing 5-methylcytosine (5mC), is probably the most extensively studied form of epigenetic regulation and will be discussed further in Chapter 4. DNAm typically acts to suppress gene expression by interfering with transcription factor binding [281]. Other mechanisms of epigenetic regulation include histone modifications, chromatin remodeling and noncoding ribonucleic acids (ncRNAs) [279].

Studies comparing DNA extracted from head and neck tumour tissue with DNA from healthy or tumour-adjacent tissue have revealed noticeable differences in the pattern of DNAm [281-284]. Cancer cells are associated with both frequent gene promoter hypermethylation, which refers to an increase in genomic 5mC within the regulatory regions of genes [285 281 286] and increased global (non-specific) hypomethylation, which describes an overall decrease in the level of 5mC across the genome. In contrast to hypermethylation, which typically results in transcriptional silencing of key genes, global hypomethylation is capable of reactivating methylation-silenced proto-oncogenes [281].

Some of the most frequently hypermethylated genes in HNC are summarised below and in Table 10. These genes cover a wide range of biological pathways, including cell cycle control (*CDKN2A*, *CDKN2B, p53*), apoptosis (*DAPK*), Wnt signalling (*APC*, *WIF1*, *RUNX3*), cell adhesion (*CDH1*), and DNA-repair (*ATM*, *MGMT* and *hMLH1*) [282]. For a more comprehensive list of aberrantly methylated genes see Gasche *et al* (2012) [281].

Hypermethylation of the *CDH1* gene, which encodes the adhesion protein E-cadherin, has been associated with HNC in multiple studies, although the reported frequency of methylation at this gene varies considerably- from as little as 7% to as high as 85%. [281 286]. For this reason, *CDH1* hypermethylation may not provide a reliable biomarker for HNC;

nonetheless, there is evidence to suggest that it may be associated with invasive tumour behaviour [287] [288]; and could therefore have prognostic significance.

CDKN2A and *CDKN2B* hypermethylation have also been associated with metastasis and overall poor survival in people with HNC [281]. The *CDKN2A* gene encodes several proteins, two of which - $p16^{INK4A}$ and the $p14^{ARF}$, are essential for regulating cell growth and division[289]. *CDKN2A* promoter hypermethylation turns off the production of these key tumour suppressor proteins, allowing cells to grow unchecked. The estimated prevalence of $p16^{INK4A}$ and $p14^{ARF}$ promoter hypermethylation in HNC varies from 12-88% and 14-44%, respectively, depending on the samples and methodologies used by different research groups [281]. *CDKN2B*, is another tumour suppressor gene that lies adjacent to *CDKN2A*. It provides instructions for making the cyclin dependent kinase inhibitor protein $p15^{INK4B}$, which controls cell cycle G1 progression [281] [290]. The existing evidence suggests that 9-28% of HNC tissues exhibit *CDKN2B* hypermethylation. Healthy or tumour-adjacent normal tissues, by comparison, typically lack *CDKN2B* methylation [281].

The DNA repair genes *MGMT* (O-6-methylguanine-DNA methyltransferase) and *MLH1* (mutL homolog 1) are also frequent targets of epigenetic modification in HNC, according to multiple independent studies (up to 75% of HNC tissues demonstrate *MGMT* and *MLH1* gene silencing via hypermethylation) [281] [291]. *MGMT* plays an important role in preventing carcinogenesis as it removes mutagenic adducts from O(6)-alkyl-guanine in DNA [281], whilst *MLH1* fixes errors that are made when DNA is copied, which helps to reduce the accumulation of mutations and maintain genomic stability.[292] Silencing of *MGMT* and *MLH1* via methylation are both believed to be an early events in HNC tumorigenesis [281] [291].

Several studies have looked for differences in the methylation profiles of HPV-positive and HPV-negative head and neck tumours. Collectively, there is consistent evidence to suggest that tumours that are associated with HPV infection exhibit a greater level of promoter methylation (2-5 times higher) compared with non-HPV associated tumors [293] [294], however, HPV-positive and HPV-negative DNAm profiles remain poorly characterised overall. One reason for this is that most studies have only evaluated a select number of genes or have focused on global markers of DNAm; few have conducted epigenome-wide analyses[293] [294]. In addition, surprisingly few methylation studies have focused solely on the oropharynx but have instead grouped multiple sites in the head and neck region together. This is surprising given that HPV infection is a much greater risk factor for OPC compared to other HNCs.

Some of the key gene families found to be differentially modulated include cell-cycle genes (*CDKN2A*, *CCNA1*), JAK-STAT pathway genes (*JAK3, STAT5A*) and cadherin family members (*CDH8*, *CDH15*, *CDH11*, *PCDH8-10*) [276] [294] [295].

Table 9: Frequency of selected genes recurrently mutated in HNC.

Gene	Frequency (%)	Ref(s)
HPV (+)		
E6/E7	100%	275 296-298
PIK3CA	22-56%	275 296-298
TP63	16-28%	296 297
TRAF3	22%	296
E2F1	19%	296
NOTCH 1/3	17%	296
FGFR3	11-14%	296 297
HLA-A/B	11%	296
EGFR	6%	296
HPV (-)		
TP53	73-84%	275 296 297
CDKN2A	25-57%	275 296 297
PIK3CA	13-34%	296 297
PIK3CB	13%	297
FADD	32%	296
FAT1	14-32%	275 296
CCND1	13-31%	275 296 297
NOTCH1/2/3	16-29%	275 296 297
TP63	19%	296
EGFR	12-15%	296 297
HNSCC		
CDKN2A	74%	296
TP53	66-79%	296 298
FAT1	46%	296
TP63	26%	296
CCND1	23%	296
MAML1	23%	296
EGFR	17%	296
TNK2	17%	296
AKT1	14%	296
SRC	14%	296
NOTCH1	14%	298

Table 10: Genes frequently hypermethylated in HNC.

Gene	Function	Ref(s).
APC	WNT signalling	279 281 282 299
ATM	DNA damage repair	279 282 300
CCNA1 (Cyclin A1)	Cell cycle regulation	281 294 300-302
CDH1 (E-cadherin)	Cell adhesion	279 281 282 299-305
CDKN2B ($p15_{INK4B}$)	Cell cycle regulation	279 281 282 300 306 307
CDKN2A ($p16_{INK4A}$/ $p14_{ARF}$)	Cell cycle regulation	279 281 282 299-301 304-310
DAPK	Apoptosis	279 281 282 300 302 303 305 306 308 310
DCC	Tumour suppressor	281 282 300 302 306 309
EDNRB	Receptor activity	281 282 309
EPCAM	Cell adhesion molecule	281 311
FHIT	Cell cycle progression	281 312
hMLH1	DNA repair	282 291 299 303 304 313
HOXA9	Cell differentiation	279 299
MGMT	DNA damage response	279 281 282 299 300 302-305 308 310
MINT1/2	Movement, exocytosis & adhesion	281 299 302 306
MINT27	Transcriptional regulator	281 306
MINT31	Transcriptional regulator	281 299 306
NPY	Cell proliferation	279 299
Notch1	Receptor activity	281 294 314
p53	Cell cycle regulation	281 282 304 307
PTEN	Differentiation, proliferation, invasion, apoptosis.	282 299

Table 10 continued.

pRB	Tumour suppressor	282 299
RARβ	Nuclear receptor	281 282 300-302 305
RASSF1	Cell cycle, apoptosis.	279 281 282 300 303 305
RECK	Protein binding/peptidase inhibitor	281 315
RUNX3	WNT signalling	281 282 299 303 310 316
SFRP family	Protein binding	281 282 310 317
TIMP3	Extracellular matrix degradation	279 282 300 302
VHL	Transcription factor binding	281 299 307
WIF1	WNT signalling	281 282 299 303

It is becoming increasingly apparent that environmental and lifestyle risk factors (e.g. smoking and alcohol drinking) are capable of instigating a range of epigenetic alterations, which could mediate at least some of the associations of these exposures with HNC risk. Smoking, which is a major risk factor for HNC, is considered one of the most significant environmental modifiers of peripheral-blood DNA [318]. It is associated with both promoter hypermethylation of tumour suppressor genes and genome-wide hypomethylation, especially in long-term tobacco users [281 319]. One of the ways in which cigarette smoke can influence DNAm profiles in people with HNC is by impairing DNA methyltransferase (*DNMT*) expression, which is responsible for catalyzing the addition of a methyl group to DNA, both at transcript and protein level (reviewed in [320]).

The ability to better characterise lifestyle exposures in people with HNC using persistent epigenetic markers will be discussed in Chapter 4.

3.7.3. Metabolomic profiles

Altered metabolism is a hallmark of cancer, being necessary to support continuous growth and proliferation [321 322]. More specifically, metabolic reprogramming enables cancer cells both to obtain and utilise potentially unconventional nutrient sources, in order to build new biomass, and influence surrounding normal cells. This can in turn have direct or indirect effects on gene expression, for instance by mediating the addition and removal of epigenetic marks from chromatin [321 323]. Pavlova *et al* (2016) [321], published a review of cancer-

associated metabolic changes, wherein they further categorised these alterations into six distinct ‘hallmarks’ of cancer metabolism; these include:

- deregulated uptake of glucose and amino acids;
- use of opportunistic modes of nutrient acquisition;
- use of glycolysis/TCA cycle intermediates for biosynthesis and NADPH production;
- increased demand for nitrogen;
- alterations in metabolite-driven gene regulation; and
- metabolic interactions with the microenvironment.

The reasons why cancer cells engage in seemingly inefficient energetic processes (i.e. aerobic glycolysis) is discussed further in Sandulache *et al* (2012) [324].

Metabolomics, which is a relatively new but rapidly emerging technology within the omics field, is becoming an increasingly utilized tool for investigating the perturbations in metabolic pathways in HNC. Metabolomics is defined as the systematic identification and quantification of all, or specific, metabolites within a biological sample (cell, tissue, blood, saliva or urine), at a specific time. The term ‘metabolite’ is typically restricted to small molecular-weight products of metabolism, typically less than 15000 Dalton (Da). There are a variety of metabolomic platforms and technologies available, but the most used techniques include mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. A detailed description of these different analytical techniques, including a comparison of their advantages and limitations of each, is provided in Chapter 4.

It is becoming increasingly apparent that people with HNC possess distinct metabolomic profiles compared with healthy controls and people with pre-cancerous legions, though studies are typically based on small samples. A summary of metabolomic-based studies on HNC is provided in Table 11. Of note, numerous studies comparing samples from people with HNC and healthy controls, point towards altered levels of several amino acids, such as alanine, glutathione, histidine, isoleucine, valine and lysine [325-328], as well enhanced glucose consumption and Lactate production [325 326 329-332], supporting the role of altered energy metabolism in HNC. In addition, elevated levels of choline-containing metabolites were also detected in HNC samples in numerous studies in (tissue and peripheral blood) [333-335]. Abnormal choline metabolism has been observed in several other cancer types and is regarded as a metabolic biomarker of tumour development and progression [271 329]

Abbreviations for Table 11: ***CE-TOF-MS****, capillary electrophoresis time-of-flight mass spectrometry;* ***CE-MS****; capillary electrophoresis-mass spectrometry;* ***GC,*** *gas chromatography;* ***HPLC****, high performance liquid chromatography;* ***HR-MAS****, high-resolution magic-angle spinning;* ***HNSCC****, head and neck squamous cell carcinoma;* ***LC-Q-TOF****, liquid chromatography quadrupole-time of flight;* ***MRS****, magnetic resonance spectroscopy;* ***MS****, mass spectrometry;* ***NHOK****, normal human oral keratinocytes;* ***NMR****, nuclear magnetic resonance spectroscopy;* ***OC****, oral cavity;* ***OLK****, oral leukoplakia;* ***OLP****, oral lichen planus;* ***OSCC****, oral squamous cell carcinoma;* ***PCA****, principle component analysis;* ***PLS-DA****, partial least squares discriminant analysis;* ***TCA****, tricarboxylic acid cycle;* ***UPLC-Q-TOF-MS****, ultraperformance liquid chromatography coupled with quadrupole/time-of-flight mass spectrometry. Metabolites:* ***Ace****, acetate;* ***Ala****, alanine;* ***Asn****, Asparagine;* ***Cho****, choline;* ***Cit****, citrate;* ***Cre****, creatine;* ***Crea****: creatinine,* ***Cys****: Cysteine;* ***GABA****, γ-aminobutyric acid;* ***Gln****, glutamine;* ***Glu****, Glutamic acid;* ***Gluc****, glucose;* ***Glut****, glutathione;* ***Glyc****, glycine;* ***His****, histidine;* ***Lac****, lactate;* ***Ileu****, isoleucine;* ***leu****, leucine;* ***Pro****, proline;* ***PUFA****, polyunsaturated fatty acid;* ***Ser****, Serine;* ***Suc****, succinate;* ***Tau****, taurine;* ***Val****, valine*

Table 11: Summary of metabolomic-based studies on HNC.

Ref.	Sample	N	Technology	Metabolomic findings
[336]	Saliva	50 HNSCC, 77 controls	HPLC	Increased Glut levels in HNSCC cases.
[337]	Saliva	20 OC, 20 OLP 7 OLK, 11 healthy controls	HPLC/MS	Metabolic profiling data distinguished between OSCC, OLP and OLK, based on PCA-approaches.
[338]	Saliva	69 OC, 87 healthy controls	CE-TOF-MS	28 metabolic traits were differentially expressed in cases compared to controls.
[339]	Saliva	37 OC, 32 OLK, 34 healthy controls	UPLC-QTOFMS	41 metabolites were differentially expressed in OSCC relative to controls and 61 metabolites measured were differentially expressed in OSCC relative to OLK. The most significant discriminant metabolites were GABA, Phe and Val, n-eicosanoic acid and lactic acid
[335]	Plasma	33 OC, 5 OLK, 28 healthy controls	NMR	Metabolic profiling data distinguished between OC, OLP and controls based on PLS-DA.
[334]	Serum	15 OSCC, 10 healthy controls	NMR	OSCC showed a distinct signature of altered energy metabolism, which included altered lipolysis (an accumulation of ketone bodies), a distorted Krebs cycle (e.g. reduced Cit, Suc, and formate), and amino acid catabolism [increased 2-hydroxbutyrate, ornithine, Asn].
[332]	Serum & tissue	25 HNSCC	GC/MS	Serum levels of several metabolites related to the glycolytic pathway, such as Gluc, were higher in patients with HNSCC. Levels of several amino acids were lower. There were differences in 9 metabolites in sera of individuals who had disease relapse compared to those who did not.
[333]	Tissue	135 HNSCC tissue, 40 normal	MRS	Tau, Cho, Glu, lactic acid, and lipid were found to have diagnostic potential.
[340]	Urine	37 OC, 32 OLK, 34 healthy controls.	GC/MS	The urinary metabolite profiles of OSCC, OLK and healthy control samples could be clearly discriminated. The most differentially expressed metabolites were hippurate, 6-hydroxynicotic acid, Ala, Tyr, Val, Ser and Cys. 6-hydroxynicotic acid and valine in combination provided the best discrimination between OSCC and healthy controls. The combination of cysteine,6-hydroxynicotic acid and tyrosine provided the best discrimination between OSCC and OLK.

Table 11 continued.

Ref.	Sample	N	Technology	Metabolomic findings
327	Tissue	19 HNSCC, 13 normal control, 3 cervical lymph nodes	MRS	Cho/Cre ratio was higher in tumor tissue than in normal tissue. Ala and Ileu were detected in 15 of 19 tumor samples and found in only one of 13 samples of normal tissue. Glut, His, and Val were found in eight, 10, and 12 tumor samples, respectively, whereas each amino acid was detected in only one of 13 normal tissue samples.
325	Tissue	159 OSCC (tumor tissues, neighbouring margins and bed tissues)	HR-MAS NMR	Malignant tissues had higher levels of Glut, Cho, phosphocholine, Lac, Ace, Tau, Glyc, Lue, lysine, Ileu and Ala, and lower levels of Cre and PUFA.
326	Tissue	22 HNSCC	HR-MAS	HNSCC tissues showed elevated levels of lactate, amino acids including Leu, Ileu, Val, Ala, Gln, glutamate, aspartate, Glyc, Phe and Tyr, Cho containing compounds, Cre, Tau, Glut, and decreased levels of triglycerides.
329	Cell lines	5 HNSCC cell lines, 3 cultures of normal human oral keratinocytes	NMR	There were alterations in the levels of several metabolites involved in multiple metabolic events, including Warburg effect, oxidative phosphorylation, energy metabolism, TCA cycle, glutaminolysis, hexosamine pathway, osmo-regulatory and antioxidant mechanisms.
341	Saliva	22 OSCC, 21 healthy controls	CE-MS	25 metabolites were identified as potential markers to discriminate between people with OSCC and healthy controls.
342	Saliva, tissue	24 OC (18 SCC tissue), 44 controls (paired tumour and control tissues).	CE-TOF-MS	85 and 45 metabolites showed significant differences between tumour and matched control samples, and between salivary samples from OC and controls, respectively. 17 metabolites showed consistent differences in both saliva and tissue-based comparisons.

Table 11 continued.

Ref.	Sample	N	Technology	Metabolomic findings
[343]	Saliva	101 OCC, 58 OPC, and 35 normal controls	NMR, LC-MS/MS, LC-Q-TOF	Glyc and Pro were different between OCC and controls. Four metabolites, including Glyc, Pro, citrulline, and ornithine were associated with early stage OCC.
[331]	Cell lines	3 HNC cell lines as well as normal fibroblasts.	NMR	Alterations in the amounts of threonine, ribose, n-acetyl, malonate, methylmalonate, malonate, threitol, n-acetyl glutamate, ethanolamine, phosphoetanolamine, unsaturated lipids (CH2) and Lac.
[330]	Tissue	32 OSCC (tumour + surrounding normal tissues)	CE-TOF-MS	Enhancement of glucose consumption and Lac production was observed in OSCC tissues. Fumarate and malate in were significantly higher in OSCC tissues compared to control.
[328]	Cells	5 HNSCC patients, NHOK from 3 donors	NMR	HNSCC cells exhibited altered levels of various metabolites related to Warburg effect, oxidative phosphorylation, energy metabolism, TCA cycle anaplerotic flux, glutaminolysis, hexosamine pathway, osmo-regulatory and antioxidant mechanism.
[344]	Serum	140 HNC patients	LC	High serum levels of methionine and alanine had a positive prognostic impact in univariate models. High serum levels of cystine predicted lower OS. Methionine remained a positive prognostic factor in multivariable models.
[345]		28 tissue samples obtained from 7 HNC cases and 7 controls + oral washes	LC/MS	Eight metabolites were elevated in tumor tissues by 1.9- to 12-fold compared to normal tissues. N-acetylputrescine was the most elevated. 2-hydroxyglutarate (2-HG, a TCA cycle analog) and glycerol-3-monophosphate (3-GMP) were only detected in tumor tissues. Levels of acylcarnitines were elevated in oral washes of HNC cases.

3.8. Survival rates after HNC diagnosis

Depending on the type of HNC, between 61% and 86% of people are expected to survive for at least a year following diagnosis (Table 12). Net survival rates fall to between 28% and 67% at five years and 19% - 59% at ten years. Salivary gland cancers have the best prognosis, with a 5-year net survival rate of around 76% (2009-2013) [346]. At the other end of the spectrum, hypopharyngeal cancers have the worst prognosis, with a 5-year net survival rate of just 28%. This discrepancy is largely a reflection of late presentation. In general, hypopharyngeal cancers remain asymptomatic until they are advanced and consequently, 80% of all cases present at stage III or IV [8].

HNC survival rates in England are similar in males and females; however, the rate for females with laryngeal cancer was not available at the time of writing this thesis due to the low number of cases [346].

For oropharyngeal cancers, the 5-year net survival rate is around 66% (Table 12). However, studies have consistently demonstrated that people with HPV-positive oropharyngeal cancers have improved overall and disease-free survival compared to their HPV-negative counterparts (presenting with a the same cancer stage) [347] [348], which is attributed in part to an improved therapeutic response [347]. Although the majority of people with HPV-positive OPC do well following treatment, a high price may be paid in terms of both acute and chronic toxicities. Because of this, there are now several on-going clinical trials looking at de-escalating treatment strategies for people with HPV-positive OPC to reduce unnecessary treatment morbidity [349-351].

Table 12: Head and neck survival by sub-site for males and females.

HNC site	Overall			Males			Females		
	1-year	5-year	10-year	1-year	5-year	10-year	1-year	5-year	10-year
Hypopharyngeal	60.5	27.8	19.1	60.4	27.2	17.9	60.7	30.2	23.3
Laryngeal	*	*	*	85.3	65.4	54.7	*	*	*
Oral cavity	78.4	56.1	45.2	77.8	53.5	42.2	79.2	59.8	49.6
Oropharyngeal	83.7	65.6	57.7	83.8	65.5	57	83.6	66.0	59.4
Salivary gland	85.8	67.0	59.3	82.5	58.3	50.7	89.8	77.8	70.3
Sinus	74.8	51.4	42.6	73.0	50.3	42.5	77.5	53.2	42.5

Based on CRUK figures [346].

3.9. Head and neck cancer prognostic factors

Prognostic factors (also called prognostic markers) are those which, in people who have already developed a condition or disease, predict the outcome or natural history of that illness [352]. Therefore, in contrast to risk factors, prognostic factors are not necessarily causally related to disease. Prognostic factors can be divided into 3 main categories: those relating to the individual, the tumour, or the environment. Individual or host-related prognostic factors include inherent demographic characteristics (e.g. age, gender, race), social and economic attributes (e.g. income, education, marital status) and clinical characteristics (e.g. comorbidity). Tumour-related factors are those that concern tumour pathology, anatomic disease extent, and tumour biology. Environment-related factors largely reflect lifestyle exposures, such as tobacco and alcohol use.

3.9.1. Individual-related HNC prognostic factors

3.9.1.1. Age

Age is frequently reported as an independent prognostic factor for overall and disease-free survival in HNC [46 353-356]. This can be explained in part by an increase in non-cancer-related mortality with increasing age [356]. Older people (usually defined as people 65+ years in HNC studies) experience greater acute and late toxicity as the intensity of HNC treatment increases [357], often because of comorbidity or impaired functional status. In particular, the addition of chemotherapy to radiation can increase toxicity and reduce tolerance to

therapy. As a result older people may not be considered suitable candidates for multimodality treatment, which could impact on prognosis [357].

3.9.1.2. Gender

Gender is less well defined as an independent prognostic factor for HNC [358-360], though some studies suggest a lower risk of death among women [361-364]. In one study, which included 238,608 HNC cases from 86 cancer registries, relative survival was 75% and 67% at 1-year and 50% and 36% at 5 years in women and men, respectively [364] , when all HNC sites except larynx were combined. The reasons for this are unclear and probably complex. However, the authors suggest that it may be explained in part by the detection of cancer at an earlier stage in women than in men, which may result from the fact that women are typically more engaged in health and consult physicians sooner [365]. The similar survival figures observed among men and women with laryngeal cancer in this study was attributed to the fatal sub-sites, particularly the supra-glottis, being more frequent in women than men (37% vs. 23%).

3.9.1.3. Race/ethnicity

There is some evidence to suggest that ethnicity is a prognostic factor for HNC survival, although most of the evidence comes from observational studies conducted in the US. In particular several studies report worse survival for African-Americans compared to White and Hispanic Americans [366-371], largely attributed to socioeconomic (SE) differences. Nevertheless, several studies demonstrate that even after adjusting for factors such as age and stage at presentation, treatment modality and health insurance status, ethnicity remains an independent prognostic factor for OS [366 369 371]. Some authors suggest that disparities may be driven in part by racial differences in HPV-associated OPCs, with HPV prevalence being lower in African Americans [372 373]. The extent to which genetic factors could be contributing to racial differences is relatively unexplored. One study reported a loss of the *CDKN2A* and a gain of *SCYA3* in African Americans [374]. *CDKN2A* has been linked to malignant progression in HNC [375 376], whilst *SCYA3* (otherwise known as macrophage inflammatory protein-1α [MIP-1-alpha]), has been implicated in tumour lymph node metastasis [377].

3.9.1.4. Socioeconomic position

Socioeconomic status (SES) affects overall survival [378-384]. Estimates of the magnitude of the difference between the most and least deprived groups vary, however [385]. This is probably because studies have used different indices of socioeconomic status (SES) and/or have

adjusted for different confounding variables [385]. SES can be measured at an individual level (e.g. educational attainment, occupation, unemployment or marital status), or by using area-based measures of deprivation. Using the income domain of the Indices of Multiple Deprivation (IMD), a recent report produced by the National Cancer Intelligence Network (NCIN) in partnership with Cancer Research UK (CRUK) showed that between 2007-2011, European age-standardised HNC mortality rates in the UK were 218-298% higher for males living in the most deprived areas compared to the least deprived, and 59-257% higher for females (for cancers of the larynx, oral cavity and oropharynx) [140 386]. One explanation for this is that people living in low SE areas are frequently diagnosed at later disease stage, possibly as a result of poorer access to health care facilities or less awareness of the signs and symptoms of oral cancer [387] . Another explanation is that the incidence of (non-cancer) smoking-and alcohol-related comorbidity is greater in deprived neighbourhoods.

3.9.1.5. Comorbidity

Comorbidity- the presence of one or more additional medical conditions - has been shown to be a strong prognostic factor for HNC survival [388-394]. In a meta-analysis of 10 studies including 22,932 people with HNC, comorbidity was found to increase overall-mortality risk by around 40% (versus no comorbidity), but it did not have an impact on cancer specific survival. Comorbidities are common in this population due to the association of HNC with tobacco and heavy alcohol use, combined with advanced age at time of diagnosis [388 394]. The literature suggests that around 60% of people with HNC have concurrent illness, of which 20% carry a severe comorbid burden [395] e.g. recent myocardial infarction, recent stroke or acute hepatic disease [396]. These figures are relatively high when compared to other cancer sites [397]. The presence of comorbid conditions is associated with higher rates of surgical and postoperative complications [398-400]. Consequently, people with comorbidities often undergo conservative and less radical procedures [395 401] which can affect HNC outcomes. Several instruments have been developed to quantify comorbidity [400]; the Adult Comorbidity Evaluation 27 (ACE 27) and Charlson Comorbidity Index (CCI) are the most frequently used indices in HNC. The ACE-27, which will be referred to again throughout this thesis, was modified from the Kaplan–Feinstein Index for use in cancer patients. The instrument grades 27 individual conditions as mild, moderate, or severe based on the severity of organ decompensation and prognostic impact and then assigns an overall comorbidity score that is equal to the highest ranked single ailment [402](further details are provided in Chapter 5). The CCI by contrast, takes into account the presence of 19 diseases; each condition is assigned a weight of 1 to 6 [403], based on the estimated 1-year mortality HR from a Cox proportional hazards model [404]; the weights are then summed to produce an overall comorbidity score

ranging from 0–37 [405] [406]. Each comorbidity instrument is associated with its own advantages and disadvantages [406] [407]. The ACE-27 is well validated for use in people with HNC but it may be less well suited to retrospective or historical studies as some of the conditions require invasive confirmation or specialist evaluation. The CCI is simple to apply, widely applied and suitable for historical studies but, as pointed out by Rotbøl Bøje *et al*, the prognostic impact of the diseases measured may have change since the instrument was first developed in the late 1980s [407].

3.9.2. Tumour-related HNC prognostic factors

3.9.2.1. Tumour stage

In addition to tumour site, tumour staging is a well-documented prognostic factor for HNC [408-412]. People with tumours that are larger and have spread to lymph nodes and other tissues in the body have poorer survival [413]. For most (but not all) HNCs, lymph node involvement automatically places the cancer at stage III (see Chapter 2). Overall, only 25%-40% of people with lymph node metastasis are expected to survive 5-years, compared to approximately 90% of people with tumours that are limited to the primary site (i.e. people with cancers that have not metastasized to the lymph nodes) [414].

In a report produced by the Oxford Cancer Intelligence Network (OCIN) looking at survival by stage, the one-year relative survival rates for early versus late stage laryngeal cancer were 97% and 76% respectively [412]. At three years, the difference in survival rates increased to almost 40% (relative survival rates of 89% for early stage disease versus 51% for late stage disease). For oral cancers, the one-year relative survival rates were 85% for early stage disease versus 72% for late stage disease and at 3-years post diagnosis, the corresponding rates were 82% and 66%. Compared to OPCs that were diagnosed with late-stage disease, OPCs diagnosed at an early stage also had higher one- and three-year relative survival rates. However, It is worth noting that, as described in Chapter 2, the staging of HPV-associated OPC has been modified to downstage the disease, for the reason that, although people tend to present with extensive disease, their prognosis is often good [415].

The location of the lymph node metastases has particular prognostic significance. For most HNCs, worse survival is observed when metastases involve lymph nodes beyond the first draining lymph node [416]. Survival is particularly poor for lymph nodes in the lower regions of the neck, i.e. levels IV and V (supraclavicular area) [417].

3.9.2.2. HPV status

People with HPV-positive oropharyngeal tumours tend to experience better overall-survival. In a retrospective analysis of participants included in the RTOG 0129 study, Ang *et al* demonstrated that, after adjustment for age, race, tumour and nodal stage, tobacco exposure, and treatment assignment, the risk of death for people with HPV-positive OPCs was 58% lower than that for people with HPV-negative tumours [418]. Here HPV status was determined using HPV-16 DNA ISH. When p16 expression was used as a proxy for HPV status, results were similar. These two methods of assessing HPV status are discussed further in the next chapter.

3.9.2.3. Genetic predictors

A number of investigators report that mutations in the gene encoding the tumour suppressor protein p53 are associated with reduced survival [419-422], and this is likely due to reduced susceptibility of cancers cells to apoptosis. However, the evidence is inconsistent [423]. HNC with mutations in *NOTCH1*, which is the second most frequently mutated gene after *TP53,* have also been shown to have a worse prognosis compared to *NOTCH1* wild-type tumours [424], and inhibition of NOTCH pathway genes has been shown to decreases cell proliferation and invasion in some studies [425]. Loss of *PTEN* expression or loss of PTEN function as a result of mutation, which are also both common occurrences in in HNC, have also been identified as potential markers of high recurrence risk [426].

3.9.2.4. Epigenetic predictors

Whilst aberrant DNAm profiles have been associated with HNC, few studies have comprehensively investigated the potential prognostic value of these biomarkers. In one study, Kostareli and colleagues [427] investigated whether the promoter methylation states of five candidate genes, including *ALDH1A2*, *OSR2*, *GATA4*, *GRIA4*, and *IRX4*, correlated with overall survival and progression-free survival in OPC. They found that a pattern of low methylation levels in *ALDH1A2 (*which encodes the aldehyde dehydrogenase family 1 member A2) and *OSR2*, and high methylation levels in *GATA4*, *GRIA4*, and *IRX4* (in tumour tissue) were associated with improved survival in 3 independent HNC cohorts (total n=220). Strikingly, the authors demonstrated that the methylation score had superior prognostic performance compared to HPV status, as measured by HPV DNA, RNA and p16 status. Larger, well-designed studies are needed to confirm these findings.

In a more recent analysis of tumour DNA samples, Zhou *et al* (2018) [283] evaluated the prognostic value of four aberrantly methylated genes (*FAM135B, ZNF610, HOXA9*, and *DCC*) and confirmed that *FAM135B* methylation was a favorable independent prognostic biomarker for overall survival in HNC (HR= 0.12; 95% CI: 0.02 to 0.69). There was no evidence of an association between the remaining candidate genes and overall survival (n= 520 HNC and 44 normal control samples).

Several studies have suggested that *p16INK4A* (*CDK2NA*) promoter methylation, which was mentioned earlier in this chapter, may serve as a useful biomarker to predict nodal metastasis and tumour recurrence in HNC [281 284]. In one study, which included 38 people with SCC of the tongue, the presence of p16 promoter hypermethylation in surgical margins was associated with a 6.3-fold increased risk of local recurrence, compared to people with negative margins [428]. p16 hypermethylation also had an unfavourable impact on DFS (HR= 2.24, 95% CI: 1.35 to 3.73) in a systematic review and meta-analysis, which included five HNC studies (n=385) [429], but the same authors found no evidence of an association between p16 methylation and OS.

3.9.2.5. Metabolic predictors

As with DNAm, there is a paucity of studies examining the influence of differentially produced metabolites on prognosis in the HNC literature, even though there is growing body of evidence to suggest that HNCs possess a distinct metabolic signature compared to controls. Yonezawa *et al* (2013) [332] analysed the sera of 17 people with HNC (largely hypopharyngeal and oral cavity cancers), using gas chromatography/mass spectrometry, and found that the serum levels of several glycolysis-related metabolites, namely glucose, ribose and fructose, were higher in individuals who experienced disease relapse (n=5) compared to those who did not. Conversely, those who had disease relapse had lower circulating levels of lysine and trans-4-hydroxy-L-proline. For oral cavity cancer specifically, the authors found that the amino acids methionine and ketoisoleucine were also lower in people who experienced cancer relapse (n=4). It is not possible to draw any conclusions from these analyses given the small sample size.

3.9.2.6. Infection biomarkers

Numerous other molecular biomarkers have been investigated as potential prognostic factors for HNC. Evidence gleaned from recent studies looking at the effects of tumour infiltrating lymphocytes on HNC outcomes suggest that neutrophil-to-lymphocyte ratio (NLR) [430-434] lymphocyte-to-monocyte ratio (LMR) [435 436], and platelet-to-lymphocyte ratio (PLR) [437]

may provide prognostic value. A systematic review of data from over 40,500 individuals in 100 studies concluded that an elevated NLR is associated with a 1.8-fold higher mortality risk (95% CI: 1.67 to 1.97) and this effect is detected across all HNC subgroups, sites, and stages. An elevated PLR also predicts poorer over-all survival, as reported in a recent meta-analysis of 9 studies (2327 individuals) [437]. Here, PLR greater than the cut-off value was associated with a 1.6-fold increase in risk of death (95% CI: 1.1 to 2.4). By contrast, pooled data from 4,260 individuals in seven cohorts (including all HNCs) demonstrated that an elevated LMR is associated with a 50% lower mortality risk (HR for OS= 0.5 [95% CI: 0.44 to 0.57]) [435]. The importance of immune function in tumour development and progression have long been acknowledged in the cancer literature [438] but at present, the exact mechanism underlying the relationships of NLR, LMR and PLR with prognosis are poorly understood.

3.9.2.7. Other molecular biomarkers

Tumour hypoxia has been associated with adverse prognosis [439]. In a meta-analysis of 28 studies, over-expression of hypoxia inducible factors (HIF 1/2α) was associated with a two-fold higher mortality risk (HR=2.12; 95% CI: 1.52 to 2.94; I2 74%) [440]. In another meta-analysis, looking at the association between tumour expression of vascular endothelial growth factor (VEGF) and survival in HNC (oral cavity 70.8% of people, pharynx 15.2%, and larynx 14%), individuals who were positive for this growth factor had almost double the risk of death at two years (relative risk [RR] = 1.88 [1.43 to 2.45]).

3.9.3. Environment-related HNC prognostic factors

3.9.3.1. Smoking

As well as being a major risk factor for HNC, smoking has consistently been associated with poorer clinical outcomes [441-452] (Table 13). The extent of its effects is poorly defined however, in part because studies have used different methodologies and measures of tobacco exposure. For instance, Duffy *et al* (2009) report a 2.4-fold higher all-cause mortality risk in current versus never-smokers [441], whilst Mayne *et al* observed an almost five-fold higher mortality risk in people with >60 pack-years of smoking compared to never-smokers [442]. Both studies were small -504 and 264 people respectively, limiting their statistical power to detect an effect. In addition to this, participants were enrolled from a single medical centre or clinical trial in the US, which effects the generalisability of their results. Historical analyses are typically larger, but they have often lack information on potentially important confounders such as socioeconomic status and comorbidities. Studies have also included different subpopulations of people, i.e. they have included different

cancer sites or tumour stages and therefore it is unclear whether the effects of tobacco exposure vary by cancer type. Moreover, few investigators have examined the potential interaction between tobacco smoking and HPV status.

3.9.3.2. Alcohol drinking

The prognostic role of alcohol consumption in HNC has yet to be fully elucidated. Table 14 provides a summary of the results of studies looking at the relationship between alcohol drinking behaviours and HNC outcomes. The evidence is conflicting. In the same studies mentioned above, Mayne *et al* (2009) reported a five-fold increased mortality risk for people who drank >35 drinks per week compared to those who abstained [442]. Duffy *et al* (2009), by contrast, found no difference in mortality risk between people with and without an alcohol problem at the time of diagnosis (based on an Alcohol Use Disorders Identification Test (AUDIT) score cut-off of ≥8 [453]) [441]. There is some evidence to suggest that heavy alcohol intake may increase the risk of second primary tumour (SPT) development. In a multicentre population-based case control study that included 876 people with laryngeal and hypopharyngeal cancers, Dikshit *et al* [448] found that people who consumed >121 grams of alcohol per day were nearly twice as likely to develop a SPT compared to people who drank 0-40 grams per day (HR=1.9; 95% CI: 1.1 to 3.2). In support of this, Do *et al* reported a risk ratio (RR) for SPT development of 1.4 for current versus never-drinkers, although with wide confidence intervals (95% CI: 0.9 to 2.2) [451].

3.9.3.3. Diet

In addition to pre-treatment smoking and drinking behaviours, dietary intake may have an influence on survival, although the evidence is limited and generally under-powered. In a small, historical study of just over 200 individuals with laryngeal cancer, Crosignani *et al* found that the consumption of vegetables, citrus fruit and olive oil was associated with a better prognosis [449]. Low fruit intake but not low vegetable intake was negatively associated with survival in another study by Duffy *et al* (HR=1.6; 95% CI: 1.1 to 2.1), however the effect attenuated on adjustment for age, marital status, education, cancer stage, and comorbidity (HR=1.3; 95% CI: 0.9 to 1.8) [441]. Similarly, in a recent analysis in H&N5000, fruit and vegetable intake were both associated with improved survival in models that adjusted for age and gender but following adjustment for health risk behaviours (smoking, alcohol drinking, fried food consumption and either fruit or vegetable intake depending on the exposure of interest), the association disappeared for fruit intake and attenuated for vegetable intake [454].

Table 13: The association of pre-treatment smoking with head and neck cancer outcomes.

Ref.	Study design	N	Site(s)	Exposure(s)	Outcome(s)	Association
[449]	Historical case control study (Italy)	213	Larynx	Cigarettes/day	OS	HR=1.3 (95% CI: 0.8, 2.0) for upper (mean 50 cigarettes/day) vs. lowest (mean 17.5 cigarettes/day) tertile.
[455]	Prospective cohort study	355	Larynx	Cigarettes/day	OS	HR=1.8 (95% CI: 1.1, 2.9) for 28+ cigarettes/day vs. 0-15 cigarettes/day.
[451]	People enrolled in a multicentre retinoid chemo-prevention trial (US)	1181	OC, pharynx, larynx	Smoking status	SPT	RR=2.2 (95% CI: 1.2, 3.5) for current vs. never smokers.
[447]	Multicentre population-based case-control study (SE Europe)	931	Larynx, HP	Cigarettes/day	OS	HR=1.8 (95% CI: 1.1, 3.1) for 26+ cigarettes/day vs. none.
[448]	Multicentre population-based case control study (SE Europe)	876	Larynx & HP	Cigarettes/day & pack years	SPT	HR=1.3 (95% CI: 0.8, 2.1) for 26 + cigarettes/day vs 0-15 cigarettes/day; HR=1.6 (95% CI: 0.8, 2.1) for 60+ pack-years of smoking vs 0-20 pack-years.
[456]	Prospective cohort study (Netherlands)	81	Tonsils	Smokers vs. non-smokers	OS	HR=5.5 (95% CI: 1.3, 23.6) for smokers vs. non-smokers.
[441]	Prospective cohort study (US)	504	Larynx, pharynx, OC	Smoking status	OS	HR=2.4 (95% CI: 1.3, 4.4) for current vs. never-smokers; HR=2.0 (95% CI: 1.2, 3.5) for former vs. never-smokers:

Table 13 continued.

Ref.	Study design	N	Site(s)	Exposure(s)	Outcome(s)	Association
[450]	Historical cohort study (Canada)	1,871	Larynx OC, HP, OP, NP	Smoking status	OS LC	HR=1.3 (95% CI: 1.2, 1.5) for former vs. never smokers; HR= 1.8 (95% CI: 1.5, 2.0) for current vs. never-smokers. HR=1.5 (95% CI: 1.2, 1.8) for current vs. former smoker.
[442]	People enrolled in an RCT of β-carotene for the prevention of secondary cancers (US)	264	OC, pharynx larynx	Smoking status Pack-years	OS	HR=2.1 (95% CI: 0.28, 16.5) for former vs. never-smokers; HR=4.91 (95% CI: 0.7, 36.0) for current vs. never-smokers. HR=5.4 (95% CI: 0.7, 40.1) for 60+ pack-years of smoking vs none.
[457]	Prospective study (US)	124	OP	Smoking status	HNC Recurrence	HR=5.2 (95% CI: 1.1, 24.4) for HPV-positive current vs. never-smokers; HR=2.9 (95% CI: 0.6,13.6) for HPV-positive former vs. never-smokers; HR=1.8 (95% CI: 0.7,4.8) for HPV-positive current vs former-smokers.
[446]	Retrospective study of people enrolled in the RTOG 9003 RTOG 0129 trials (US)	506	OP	Smoking status Pack-years	OS	HR RGOT 9003= 3.9 (95% CI:2.0, 7.5) for current vs never-smokers. HR RGOT 0129: 2.0 (95% CI:1.0, 4.3) for current vs never-smokers. HR RTOG 9003= 2.1 (95% CI: 1.4, 3.3) for >10 vs. ≤ 10 pack-years of smoking. HR RTOG 0129: 1.8 (95% CI: 1.1, 3.0) for >10 vs. ≤ 10 pack-years of smoking.

Table 13 continued.

Ref.	Study design	N	Site(s)	Exposure(s)	Outcome(s)	Association
[458]	Historical cohort study (US)	132	OP	Smoking status	Distant metastases	HR=12.7 (95% CI: 3.5–46.0) for active vs. non-active smokers.
[445]	Prospective cohort study (US)	382	Not stated	Smoking status	OS	HR male=1.6 (95% CI: 1.0, 2.6) and HR female=1.0 (95% CI: 0.5, 1.7) for current vs. never-smokers; HR male=1.7 (95% CI:1.2, 2.) and HR female=1.0 (95% CI: 0.5,1.8) for current vs. former-smokers; HR males= 2.1 (95% CI: 1.4, 3.1) and HR females=0.5 (95% CI: 0.,3,0.9) for current smoker vs. recent quitter.
[444]	Prospective cohort study (US)	89	All sites	Smoking status	Post-operative complications	Compared with never smokers, former and current smokers had complication rate ratios of 6.1 and 6.3 respectively.
[443]	Historical population-based study (Ireland)	5,652	OC, pharynx, larynx	Smoking status	DSS	HR=1.4 (95% CI:1.2, 1.5) for current vs never-smokers; HR=1.1 (95% CI: 1.0,1.3) for former vs never-smokers.
[459]	Prospective cohort study (US)	687	OC, OP, larynx, HP	Smoking status	OS	HR=2.1 (95% CI: 1.3, 3.3) for current vs never-smokers; HR=2.1 (95% CI: 1.3, 3.4) for former vs. never-smoker.
					RFS	HR=1.87 (95% CI:1.1, 3.3) for current vs never-smokers*; HR=2.2 (95% CI: 1.2, 3.9) for former vs. never-smokers.
				Pack-years	OS, RFS	Worse survival associated with every ten- year increase in pack-years in univariable models.

*Abbreviations: **C,**: confidence interval; **DSS**, disease-specific survival; **HR**, hazard ration; **LC**, local control; **N**, sample number; **OC**, oral cavity; **O,**: oropharynx; **OS**, overall survival; **HP**, hypopharynx; **Ref**, reference; **RFS**, recurrence free survival; **RR**, relative risk; **SPT**, second primary tumour; **RTOG**, Radiation Therapy Oncology Group; **SE**, South-east; **US**, united states*.

Table 14: Association of pre-treatment alcohol consumption with head and neck cancer outcomes.

Ref.	Study design	N	site(s)	Exposure(s)	Outcome(s)	Results
[449]	Historical population-based case control study (Italy)	213	Larynx	g/day	OS	HR=1.12 (95% CI: 0.7, 1.8) for upper tertile of alcohol consumption (mean=139.8 g/day) vs. lower tertile (mean=22.9 g/day).
[455]	Prospective cohort study	355	Larynx	g/day	OS	HR=1.1 (95% CI: 0.7, 1.9) for people who consumed ≥121g/day of alcohol vs. people who consumed 0-40 g/day.
[451]	Participants enrolled in a multicentre HNC retinoid chemoprevention trial (US)	1181	OC, pharynx, larynx	Alcohol consumption status	SPT	RR=1.4 (95% CI: 0.9, 2.2) for current drinkers vs. never-drinkers; RR=1.2 (95% CI: 0.8,1.9) for former drinker vs. never-drinkers.
[447]	Multicentre population-based case control study (SE Europe)	931	Larynx, HP	g/day	OS	HR=1.3 (95% CI: 1.0, 1.6) for people who consumed ≥121g/day vs. people who consumed 0-40 g/day; HR=1.2 (95% CI: 0.9, 1.5) for people who consumed 81-120 g/day vs. 0-40 g/day; HR=0.9 (95% CI: 0.8,1.2) for people who drank 41-80 g/day vs. 0-40 g/day.
[441]	Prospective cohort study (US)	504	Larynx, pharynx, OC	Pre-treatment alcohol problem	OS	HR=1.32 (95% CI: 0.9,1.9) for people with a pre-treatment alcohol problem vs. people without.

Table 14 continued.

Ref.	Study design	N	site(s)	Exposure(s)	Outcome(s)	Results
448	Multicentre population-based case control study (SE Europe)	876	Larynx, HP	g/day	SPT	HR=1.9 (95% CI: 1.1, 3.2) for people who consumed ≥121g/day vs. people who consumed 0-40 g/day; HR=1.6 (95% CI: 0.9, 2.7) for people who consumed 81-120 g/day vs. 0-40 g/day; HR=1.4 (95% CI: 0.9, 2.4) for people who drank 41-80 g/day vs. 0-40 g/day.
450	Historical cohort study (Canada)	1,871	larynx, OC, HP, OP, NP	Alcohol consumption status	OS LC	HR=1.3 (95% CI: 1.2,1.4) for active drinker vs. never-drinkers; HR=1.1 (95% CI: 1.0, 1.2) for former drinker vs. never-drinkers. HR=1.3 (95% CI: 1.2, 1.5) for active drinker vs. never-drinkers; HR=1.2 (95% CI: 1.0, 1.3) for former drinker vs. never-drinkers.
442	Participants enrolled in an RCT of β-carotene for the prevention of SPTs (US).	264	OC, pharynx, larynx	Drinks/week	OS	HR=4.87 (95% CI: 1.5, 16.3) for >35 drinks/week vs. none; HR=2.4 (95% CI: 0.6,9.3) for 22-35 drinks/week vs. none; HR=1.4 (95% CI: 0.4,5.4) for 8-21 drinks/week vs. none; HR=1.5 (95% CI: 0.4,6.1) for 1-7 drinks/week vs. none.

*Abbreviations: **CI**, confidence interval; **DSS**; **g/day**, grams of alcohol per day; **HR**, hazard ration; **LC**, local control; **N**, sample number; **OC**, oral cavity; **OP**, oropharynx; **OS**, overall survival; **HP**, hypopharynx; **Ref**, reference; **RR**, relative risk; **SPT**, second primary tumour; **SE**, South-east; **US**, united states*.

3.9.4. HNC prognostic models

Prognostication is a fundamental part of medicine. From the clinician's point of view, attaining an accurate prognosis for their patient is important as it has a decisive impact on treatment decisions (e.g. curative treatments versus palliative care). From the point of view of the individual with cancer, evidence suggests that people often find it easier to cope with a cancer diagnosis if they know what their expected prognosis is; it can give them time to make important decisions such as whether they want treatment, how best to manage treatment side effects, and how to take care of financial, family, and legal matters, amongst other things [460-462]. Making an accurate HNC prognosis can be very challenging even for the most experienced physician, as disease outcomes are based on the presence and interaction of multiple factors [463], as evidenced above. The relative contribution of these individual prognostic factors varies. Prognostic models, also referred to as clinical prediction models or clinical prediction rules, are statistical equations that predict an individual's risk of a future outcome (typically death or recurrence of disease) within a specific period of time, based on the combination of values of multiple predictors (e.g. age, gender, biomarkers) [464]. There are several papers describing the methods of model development [465 466], therefore they will not be described in detail here, but in brief, prognostic models are usually developed using multivariable regression techniques such as logistic models and survival analysis models.

There are three main phases in prognostic model research, as outlined in the Prognosis Research Strategy (PROGRESS) framework [464]:

1) Model development, including internal validation;
2) External validation;
3) Evaluation of model impact in clinical practice (e.g. influence on decision making, individual outcomes, and costs).

Most publications describe the model development phase, with few models being implemented in clinical practice [467]. Models are more likely to perform well, and therefore have clinical utility, if they are developed using large, high quality datasets, are based on a pre-defined study protocol and are validated in independent datasets selected from different settings [468]. The final point is important because the predictive performance of a model estimated on a training dataset, or development data, is frequently optimistic, owing to issues of multiple testing in a limited sample size [464]. This impacts on the generalisation of prognostic models. Another important point is that model performance may diminish over time, perhaps due to improvements in making diagnoses earlier or advances in treatments. It is recommended that researchers should first consider whether it is

possible to improve or augment existing models, either by recalibrating them (i.e. adjusting the intercept of the model and/or the relative weights of the predictors) or by adding additional predictors such as novel biomarkers or results from new imaging techniques, before developing new models [464] [468]. However, the independent effects of new prognostic markers need to be quite large in order to achieve a clinically meaningful improvement compared to standard models, which generally include the most important predictors [464] (e.g. age, tumour stage and HPV status in the case of HNC).

Table 15 summarises the results of a scoping review of HNC prognostic models in Ovid Medline. The purpose of which was to determine the size and nature of the evidence base in this area. Bibliographic searches incorporated the following search terms.

1. predict*.m_titl.
2. prognos*.m_titl.
3. decision.m_titl.
4. decide.m_titl.
5. deciding.m_titl.
6. 1 or 2 or 3 or 4 or 5
7. calculat*.m_titl.
8. mode" .m_titl.
9. nomogram.m_titl.
10. index.m_titl.
11. indices.m_titl.
12. tool.m_titl.
13. program.m_titl.
14. 7 or 8 or 9 or 10 or 11 or 12 or 13
15. 6 and 14
16. oral.m_titl.
17. oropharyn*.m_titl.
18. hypopharynx*.m_titl.
19. pharyn*.m_titl.
20. laryn*.m_titl.
21. nasopharyn*.m_titl.
22. mouth.m_titl.
23. nose.m_titl
24. nasal.m_titl.
25. head.m_titl

26. neck.m_titl.
27. paranasal.m_titl.
28. 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27
29. 15 and 28
30. Limit 29 cancer

A manual search in the bibliographies of selected articles was also conducted. Only studies that meet the following criteria are included:

1) The outcome of interest is survival (overall, disease-specific or recurrence-free);
2) Only prognostic factors that are available at the time of diagnosis were considered for inclusion in the model;
3) Models are multivariable (i.e. multiple biological/clinical/ lifestyle factors were considered for inclusion in the model).

Of the fourteen models identified, six were applicable to multiple tumour sites, three were specific to OPC, one to HPV-negative OSCCs (tongue and oral cavity), one to oral cavity cancer, one to laryngeal cancer and one to tongue cancer. The most frequently occurring predictors were age and tumour stage, which were included in eleven and twelve models, respectively. The next most common predictors were comorbidity (five models), tumour grade (three models) and cigarette smoking (five models), which was defined by number of pack years of smoking or smoking status (e.g. current/former/never).

Some of the authors derived and integrated prognostic scores into their prediction models, which were based on either gene expression profiles or clinicopathological features of the tumour. Mes *et al* (2017) for example, built a multivariable genomic model that included a 40-gene overall survival signature, in addition to established clinical and pathological prognostic variables [469]. By itself, the overall-survival signature only modestly predicted survival with an AUC of 0.63 (95% CI: (0.57-0.68); when combined with clinical and pathological variables (age at diagnosis, smoking, pTNM and a composite variable that was scored positive if extra capsular spread or tumour-positive margins or multiple metastatic lymph nodes were present), the prognostic accuracy of the model increased, yielding an AUC of 0.74 (95% CI: 0.69-0.79). This compares to an AUC of 0.51 (95% CI: (0.47-0.57) for standard pTNM staging only and an AUC of 0.66 (95% CI: 0.59-0.73) for a model that included pTNM, age and pack-years of smoking.

Almangush *et al* (2015), included a tumour budding-depth of invasion score in their model. Tumour budding was defined in this analysis as “as the presence of single cancer cells or small clusters of

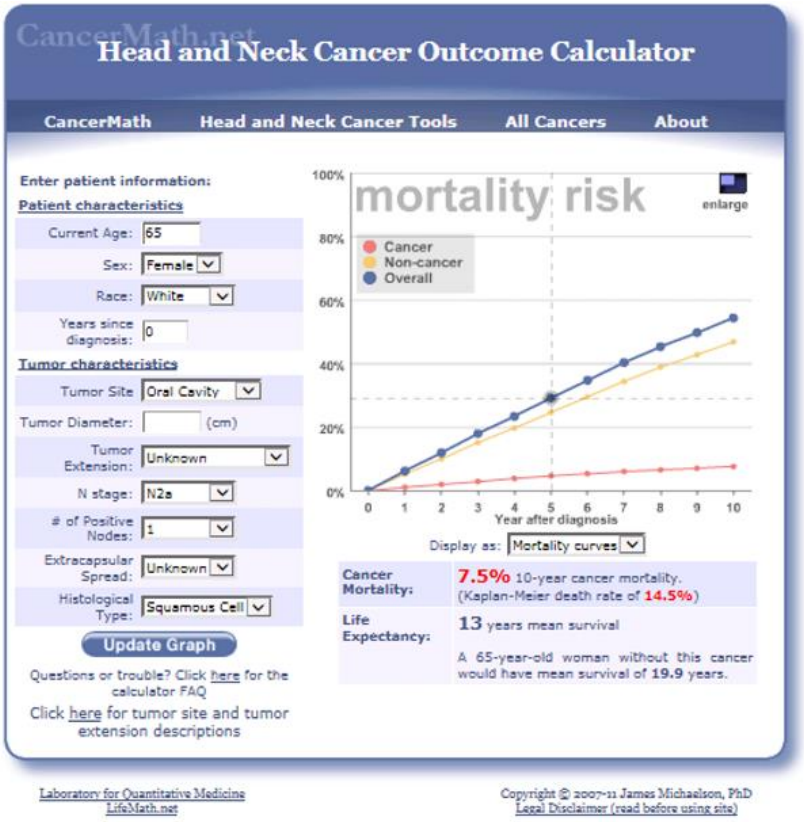
cells (<5 cells) at the invasive front" Individuals were scored "0" if neither budding nor depth of invasion were higher than a pre-defined cut-off value , "1" if only one of the parameters was higher than the cut-off, or "2" if both parameters were higher than the cut-off [470]. On multivariate analysis, a high-risk score (BD score 2) was strongly associated with cancer-specific mortality (HR= 5.11; 95% CI: 2.05 to 12.75) in early stage oral tongue squamous cell carcinoma (OTSCC). The authors suggest that the novel tumour-budding model provides a promising prognostic tool that may enable clinicians to identify people with aggressive cases of early stage disease who could benefit from multimodality treatment. In order for tumour budding scores to be introduced into routine pathology reporting, however, standardized studies including different HNCs are needed, both to identify the optimal cut-off point for characterising tumour budding in each cancer type, and to establish the best scoring method to use (e.g. which areas of the invasive front to analyse).

Two of the prediction models described in Table 15 have been converted to online tools [471] [472], which are intended to assist (not replace) clinicians with their prediction of an individual's future mortality risk. A screen shot of the first web-based calculator, which was built by Emerick *et al*. is depicted in Figure 18. The underlying model was developed using data from over 50,000 people included in the SEER database and validated externally in a dataset containing around 13,000 individuals entered into the Massachusetts General Hospital tumour registry between 1980 and 2009. The user inputs the individual's age, gender, race, tumour site, tumour dimension, tumour extension, N-stage, number of positive lymph nodes, extracapsular spread and histological type (i.e. squamous cell), by using either a drop-down menu or by entering the value manually. The calculator then predicts survival (risk of death from HNC, risk of death from non-cancer causes and overall risk of death) for each of the first ten years after diagnosis, alongside estimates of the 10-year Kaplan-Meier cancer-specific death rate and the impact of disease on life expectancy. The authors were unable to incorporate HPV data into their model due to lack of HPV-related data and cite this as a limitation. A positive feature of the online tool is that the survival information can be viewed in a number of different formats, including survival curves, death curves, pie charts or "smiley face" charts, making it a useful tool for clinicians to communicate prognostic information to their patients.

The second on-line tool, which was developed by Velazquez *et al* (2014) [472], was fitted on data from a sample of 168 people with OPC and validated in an external dataset of 189 people. The factors evaluated for their prognostic potential included age, gender, HPV status, tumour site, TNM classification, comorbidity (measured using the ACE-27), pre-treatment haemoglobin levels and smoking (pack-years) and alcohol history. For overall-survival specifically, the factors that remained in the model (and which were integrated into the on-line tool), were male gender, low pre-treatment haemoglobin levels (<median), higher T-stage, N2b-N3 stage, negative HPV status and high

comorbidity (moderate to severe). Prediction of overall-survival using the model yielded concordance statistics (C-index) of 0.82 (95% CI: 0.76 to 0.88) and 0.73 (95% CI: 0.66 to 0.79) in the training and validation datasets, respectively, indicating good (but not optimal) calibration. As a comparison, the C-index for TNM staging alone was 0.66 (95% CI: 0.61 to 0.75) and HPV alone 0.68 (95% CI: 0.61 to 0.72). Based on the values entered for the individual, the output of the on-line calculator includes estimates of two- and five- year risk probabilities for overall and progression-free survival. This information is presented visually as a histogram, and an accompanying summary of the information is provided in text beneath. The bars of the graph appear as green, orange or red, depending on whether the individual is classed as being at 'low', 'medium' or 'high' risk of death.

Figure 18: Image of web-based calculator developed by Emerick et al.



Source: http://www.lifemath.net

The on-line tool is straightforward to use and could provide a useful aid in clinical decision-making. However, it is important to note that, whilst the model was validated in an independent cohort, the samples number was small (and potentially not very generalisable), therefore further validation studies are needed.

Table 15: Summary description of HNC prognostic models described in the literature.

Ref.	Cancer site(s)	N	Population	Outcome(s)	Final predictors	Tool	Externally validated
[470]	Tongue	311	Hospital-recruited cases (Finland and Brazil)	DSS	Age, gender, tumour grade, TNM stage, tumour budding and depth of invasion (BD) score.	Equation	N
[473]	OC, pharynx, larynx	930	Hospital-recruited cases (Netherlands)	OS	Site of the primary tumour, age at diagnosis, gender, T-, N-, and M-stage, and prior malignancies.	Equation	Y
[474]	OSCC	192	Participants included in the TANRIC[1] dataset (worldwide)	Good vs poor OS	Three lncRNA panel (TN1-AS1, LINC00460 and RPS-894A10.6) and tumour grade.	Equation	N
[475]	OSSC	446	Hospital-recruited cases (China)	OS	TNM stage, tumour grade, ill-fitting dentures, oral hygiene and cigarette smoking.	Equation	Y
[476]	Lip, OC, OP, NP, HP, larynx	1282	Hospital-recruited cases (Netherlands)	OS	Age, sex, tumour location, T-, N-, M- stage, prior malignancy, comorbidity ACE-27 comorbidity score.	Equation	Y [477]
[471]	All HNCs	50,145	Participants included in the SEER dataset (US)	OS, DSS	Tumour size (T), no. of positive lymph nodes, anatomic site, tumour extension, N-stage, extracapsular spread, age and race.	Web-based calculator[2]	Y
[469]	HPV-negative OSCC	125	Hospital-recruited cases (Germany and US)	OS	40 gene overall survival signature, age, pack-years of smoking, pathological composite variable (including extracapsular spread, tumour positive margins and multiple metastatic lymph nodes).	Equation	Y

Table 15 continued.

Ref.	Cancer site(s)	N	Population	Outcome(s)	Final predictors	Tool	Externally validated
[478]	OPC	493	Participants enrolled in the NRG Oncology RTOG 0129 and 0552 trials	OS PFS	Age, pack-years of smoking, Zubrod performance status, education, anaemia, tumour p16 status, T-stage, N-stage. Age, Zubrod performance status, p16 status, weight loss, education, marital status, pack years of smoking, T-stage and N-stage.	Nomogram	Y
[479]	OC	1617	Hospital-recruited cases (US)	OS DSS	Age, race, tobacco status (never/ever smoker), comorbidity (defined using the WUHNCI) tumour diameter and clinical N-stage. Tumour size, nodal status, subsite and bone metastasis.	Nomogram	N
[480]	OC, OP, HP, larynx	1010	Hospital-recruited cases (US)	OS	Age, alcohol use, symptom severity stage, comorbidity (based on KFI) and TNM stage.	Equation	N
[481]	Larynx	788	Hospital-recruited cases (Netherlands)	OS	Age, cTNM stage, ACE-27 comorbidity score, BMI and anaemia.	Equation	N
[482]	OP	723	Hospital-recruited cases (Netherlands)	OS, PFS	Age, gender, comorbidity, pack years, T-stage, N-stage and HPV status. Age, gender, comorbidity, T-stage, N-stage and HPV status.	Equation	Y [483]

Table 15 continued.

Ref.	Cancer site(s)	N	Population	Outcome(s)	Final predictors	Tool	Externally validated
[484]	TSCC, BOTSCC	197	Hospital recruited cases (Sweden)	3-year D/R	Age, stage, diagnosis (TSCC or BOTSCC), HLA Class I expression, and CD8+TIL counts.	Equation	Y
[472]	OP	168	Hospital recruited cases (Netherlands)	OS PFS	HPV status, comorbidity, T- and N-classification, pack years of smoking, gender and pre-treatment haemoglobin levels. Gender, comorbidity, T-stage, N-stage, HPV status.	Nomogram and on-line risk calculator[4]	Y

Abbreviations: ***BSCC****, basaloid squamous cell carcinoma;* ***BOTSCC****, base of tongue squamous cell carcinoma;* ***cTNM****, clinical tumour-node-metastasis;* ***CSD****, cause-specific death;* ***DFS****, disease-free survival; D/R: death or relapse;* ***DSS****, disease specific survival; HNC, head and neck cancer;* ***HP****, hypopharyngeal;* ***HNSCC****, head and neck squamous cell carcinoma;* ***lncRNA****, Long non-coding RNAs;* ***KFI****, Kaplan-Feinstein index;* ***m****, months;* ***OC****, oral cavity;* ***OP****, oropharyngeal;* ***OS****, overall survival;* ***OSCC****, oral squamous cell carcinoma;* ***PFS****, progression free survival;* ***RTOG****: Radiation Therapy Oncology Group;* ***SCCOC****; squamous cell carcinoma of the oral cavity;* ***SUVmax****, maximum standardised uptake value;* ***TILs****: tumour infiltrating lymphocytes;* ***TSCC****, tongue squamous cell carcinoma;* ***WUHNCI****, Washington University Head and Neck Comorbidity Index;* ***yr****, years.[1] The Atlas of ncRNA in Cancer (TANRIC)* [2] *https://ibl.mdanderson.org/tanric/_design/basic/query.html.* [3] *http://www.lifemath.net/cancer/headneck/outcome/index.php* [4] *www.predictcancer.org*

3.10. Summary

HNC is a global health burden. Most incident cases occur among older individuals and among males, though the disease also represents a significant cause of morbidity and mortality among women. Development of HNC is a multifactorial process with a wide variety of individual and social risk factors, including diet, genetics and environmental exposures. The most common risk factors overall are tobacco and alcohol use, which together account for around two-thirds of all cases. Accordingly, HNC is especially common in regions of the world where tobacco and alcohol consumption rates are high. Polymorphisms in tobacco and alcohol metabolizing genes such as CYP gene family, glutathione S-transferases and ALDH may explain differences in people's risk for developing tobacco and alcohol-related HNCs, since not everyone who smokes, and drinks goes on to develop the disease.

HPV infection is another well-established risk factor for a sub-set of OPCs. HPV-positive OPCs are particularly prevalent in economically developed countries, where incidence rates have been increasing over the last few decades. Although HPV-positive tumours are often detected at an advanced stage, they tend have a better prognosis.

The recognition that HNC is not a single disease entity but rather a group of distinct cancer types, each of which may have different and interacting etiologies and clinical outcomes, has led to the development of several multifactorial prognostic models. The goal of prognostic research, which relates to the current movement towards stratified medicine, is to provide individualized outcome prediction. A review of the literature suggests that integrating clinical, molecular and histopathological variables into prognostic models could facilitate more accurate prediction. The contribution of genetic, proteomic and metabolomic measures to HNC outcome prediction is a particularly active area of research at present, but so far, no such markers have been adopted in clinical practice. Age, cancer site and stage remain the backbone of clinical prediction

Chapter 4: Capturing exposures to biological, environmental and lifestyle risk factors

4.1. Introduction

An “exposure”, in an epidemiological context, is any factor that may be associated with an outcome of interest [485]. This includes the primary explanatory variable as well as any additional variables that may be related to the outcome, for example confounders or effect modifiers, which should be taken into consideration in the statistical model. An important concern in any study design, therefore, is how to identify and characterise exposure to a given factor.

The previous chapter identified the main biological and lifestyle exposures related to HNC risk and prognosis, including tobacco and alcohol use, HPV infection, BMI and socioeconomic status. This chapter addresses the issue of how these exposures can be measured. It starts by looking at some of the instruments that are available for assessing health-related behaviours and other characteristics linked to healthy life expectancy such as biological age; it then goes on to explore different methods for detecting HPV infection, which is particularly significant for studies that include OPC. The chapter ends by highlighting the growing potential of metabolomics for identifying intermediate biomarkers linking exposure and disease outcomes.

4.2. Self-reported phenotypes

Information on health-related behaviours and socioeconomic position are commonly obtained via self-administered questionnaires. This may be the only method available to researcher. It is relatively inexpensive, easy to perform and it provides a means of assessing exposures in a way that is non-invasive and generally acceptable to respondents. As such, self-report can facilitate research that may otherwise be impossible to carry out. However, in order to obtain accurate information, respondents must have good insight into their own behaviours, must answer the questions honestly and understand what the questions are asking. Even small changes in the way a question is worded, the order the questions appear or the format of the questions (e.g. open-ended or fixed response) could result in different responses. The effects of mode of questionnaire administration and design on data quality

falls outside the scope of this thesis, however an excellent summary is provided in Bowling, (2005) [486].

There are many reasons why individuals may provide biased estimates of their own behaviours. Often, they may answer in a way that makes them 'look good', even if the questionnaire is anonymous, a phenomenon known as social desirability bias [487]. Here, a person's answer is determined by what they perceive as being socially acceptable, rather than what is true. This can result in under-reporting of unhealthy or undesirable behaviours, like smoking and alcohol drinking, and over-reporting of healthy or desirable behaviours, such as exercise. That said, self-reported smoking has been validated in numerous populations and in general, high agreement has been found between self-report and measurements of cotinine, a nicotine metabolite that is often used as an index of smoking status [488] [489]. Self-reported alcohol intake is less precise, with studies suggesting that heavy drinkers are more likely to under-estimate their alcohol intake compared to light or moderate drinkers [489] [490].

Assessments of tobacco exposure are largely based on retrospective reports of the average number of cigarettes smoked per day, which is often converted into 'pack-years' of smoking. This measure has been shown to be variable in terms of reliability and validity, largely due to individual differences in the way in which people smoke. Variations in smoking habits (e.g. frequency of puffs, puff volume and puff duration), which is termed smoking topography [491], as well as differences in the brand of cigarette smoked can lead to different levels of exposure, even in individuals who smoke the same number of cigarettes. In addition to this, the average number of cigarettes smoked per day may change over a person's lifetime. Self-report questionnaires may not fully capture cumulative life-time exposure, particularly if the respondent is unable to recall past behaviours.

Self-reported alcohol consumption is measured in a number of ways and there is an extensive literature comparing the advantages and disadvantages of each approach [492] [493]. The quantity/frequency (QF) method, which focus on 'typical' intake, and the graduated frequency (GF) approaches [494], which enquires about the frequency of drinking days for graded amounts of alcohol (e.g. the number of occasions when one or two drinks, three to four drinks, etc. were consumed), are among the most commonly utilised methods. Evidence suggests that GF produces higher estimates of alcohol intake than QF because of its ability to measure day-to-day drinking variability but QF is quick and simple to complete [495], which is important if response fatigue is an issue. Both approaches rely on recall of actual drinking episodes, which may be a particular issue for problem drinkers.

4.3. Biochemical measures of tobacco and alcohol exposure

As alluded to above, chemical biomarkers of tobacco and alcohol intake such as serum or urinary cotinine can provide an objective measure of exposure, including passive exposure in the case of smoking, as they do not rely on valid self-report and are not therefore vulnerable to issues of inaccurate recall or social desirability bias. However, the "window of assessment", that is the amount of time that the marker continues to be positive following exposure to tobacco or alcohol, is often limited [496]. The half-life of cotinine for example, i.e. the time required for the concentration in the body to decrease by half, ranges from between 2-3 days (possibly more if it is detected in hair); As such, cotinine can only provide a measure of recent tobacco exposure. Similarly, the physical presence of ethanol in the body is short-lived. Several indirect biomarkers of (predominantly heavy) alcohol consumption have been identified (Table 16), including Gamma–glutamyltransferase (GGT), Ethyl glucuronide (EtG) and the ratio of Hydroxytryptophol (5–HTOL) to 5–hydroxyindole–3–acetic acid (5–HIAA), but even these can only characterise drinking patterns over the last few weeks, and hence, are better suited for use in alcohol treatment studies i.e. to evaluate treatment efficiency. For studies which aim to assess the health risks accumulated over time, biomarkers of long-term tobacco and alcohol exposure are needed.

4.4. Epigenetic predictors of health and lifestyle

As stated in Chapter 3, epigenetic mechanisms are modifications that affect the activity of the DNA without altering the DNA sequence itself, as indicated by the prefix "epi-" which literally means "on top of" or "over" [497] [498]. DNA methylation, which is described in detail below, is by far the most frequently studied form of epigenetic modification. Studies of the temporal stability of DNAm indicate that it typically very stable over time. Indeed, past smoking has been associated with DNAm levels decades after cessation [499]. In this way, the epigenome can act as an "historical archive" of past exposure to cancer risk factors [500], which could be important in in the context of epidemiological studies.

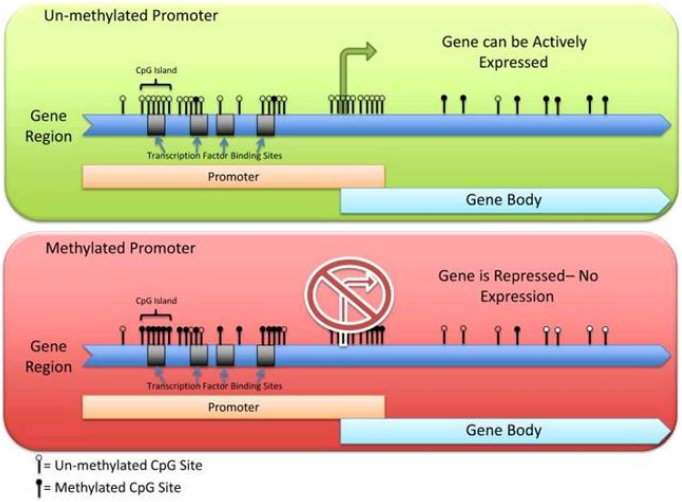
4.4.1. What is DNA methylation?

DNAm involves the addition of a methyl group (CH_3) to the fifth carbon atom of a cytosine base, producing 5-methylcytosine (5mC) (Figure 19). The chemical addition of the methyl

group at the cytosine base is made by a covalent bond, which, as highlighted above, is relatively stable over time. Methylation typically occurs at cytosine–phosphate–guanine (CpG) sites, which are regions of the DNA where a cytosine nucleotide (C) precedes a guanine (G) nucleotide in a 3-5' linear sequence [501], resulting in two methylated Cs situated diagonally from each other on opposite strands of the DNA. CpG sites are predominantly concentrated in the promoter regions of genes, which is where the transcription of DNA to RNA is initiated. When CpGs in the promoter region of a gene are methylated, the gene is typically inactive, i.e. transcription is "dimmed down" or turned off. As such, DNAm plays a vital role in several important biological processes such as embryonic development, genomic imprinting and X-chromosome inactivation[502]. Whilst DNAm marks are considered relatively stable over time, they are often modifiable, a feature that is important for the regulation of developmental processes [500].

Multiple techniques for DNA methylation analysis, including bisulfite sequencing, pyrosequencing and methylation-specific polymerase chain reaction (PCR) and have developed over the past couple of decades. The different analytic techniques have been reviewed in depth elsewhere [503] [504]. This thesis focuses on array-based approaches, which are a popular choice because they facilitate simultaneous analysis of a large number of samples and are relatively cost-effective

Figure 19: Regulation of transcription by DNAm



Top panel: *unmethylated CpG dinucleotides (open circles) allow binding of transcription factors (TF) to gene regulatory regions and gene expression is activate.* ***Bottom panel:*** *methylated CpGs (filled circles) block TF binding, thereby inhibiting gene expression. Image source* [497]

Table 16: A summary of some of the characteristics of available alcohol biomarkers, Adapted from [496] and [505].

Biomarker	Type of drinking characterised	Advantages	Disadvantages
Traditional biomarkers			
Gamma-glutamyltransferase (GTT)	Continuous, rather than episodic, heavy drinking (~ 70 drinks/wk for several weeks). Returns to normal levels with 6 weeks of abstinence.	Inexpensive, widely available. Used clinically.	Not very sensitive (61%) [506]. Elevated GGT levels may also arise as a result of hepatobiliary disorders, obesity, diabetes, hypertension, and hypertriglyceridemia. Large numbers of false negatives.
Aspartate aminotransferase (ASAT)	Heavy drinking (undefined). Returns to normal levels within 7 days, but considerable variability in declines with abstinence.	Can be analysed by a relatively simple immunochemical procedure, for which antibodies against ASAT are commercially available.	Several sources of false positives.
Alanine aminotransferase (ALAT)	Heavy drinking (undefined). Unknown half-life.	Used clinically.	Less sensitive than AST. Several sources of false positives.
Macrocytic volume	Heavy drinking (undefined). Half–life of ~ 40 days.	The testing methodology is easy and inexpensive.	Low sensitivity as a stand-alone biomarker.
Carbohydrate-deficient transferrin (CDT).	60+ g/d for at least 2 weeks. Normalise with a mean half–life of 2–4 weeks of abstinence.	Low rates of false positives. Kits to isolate and quantitate CDT in serum are commercially available.	Difficult to measure accurately. Relatively high rate of false negative results. Women tend to have higher CDT levels than men [506].

Table 16 continued.

Biomarker	Type of drinking characterised	Advantages	Disadvantages
Emerging biomarkers			
Hexosaminidase (Hex)	At least 10 days of drinking > 60g/d. Serum levels return to normal after 7–10 days of abstinence whilst urine levels normalise after 4 weeks.	Specificity of 96 % for serum and urine markers combined.	Elevated serum levels can occur with liver diseases, hypertension, diabetes mellitus, silicosis, myocardial infarction & thyrotoxicosis.
Sialic acid (SA)	Positive relationship between alcohol intake and SA levels in serum.	Can be measured in saliva.	The dose of alcohol needed to increase SA is undefined. Levels also rise in conditions other than heavy drinking, including people suffering from tumours, inflammatory conditions, diabetes, & cardiovascular disease.
5-HTOL/5-HIAA	Recent consumption of even quite low levels of alcohol. Returns to normal levels after 3–4 days (half-life 2–3 hr).	The response of 5–HTOL to alcohol is dose dependent.	The role of confounders such as age, sex, or concomitant diseases are unclear.
Ethyl glucuronide (EtG)	Best detects heavy versus light drinking over five days.	Present in various body fluids and hair. Age, sex, ethnicity, and severity of liver disease have no influence. Very specific for alcohol.	Investigations of EtG are preliminary in nature. No commercial kits have been marketed. Bacterial degradation possible (urine).
Fatty acid ethyl esters (FAEE)	Recent, heavy alcohol use [506].	Up to 100% sensitivity, 90% specificity reported [506].	Fat production may depend on age/sex/hormones.

4.4.2. DNAm-based predictors of smoking, alcohol consumption, educational attainment and BMI

There are multiple examples of the utility of DNAm in trait prediction in the literature. Methylation levels at a single loci (cg05575921) within the aryl hydrocarbon repressor gene (AHRR) for example, has been shown to provide a sensitive and specific biomarker for cigarette consumption, yielding an AUC of 0.99 [507]. The effects of smoking on DNAm at this site are unaffected by alcohol drinking, meaning that smoking and alcohol consumption - behaviours that frequently co-occur, can be estimated simultaneously [508].

Cigarette smoking has in fact been found to have a broad impact on the genome. In a large meta-analysis of genome-wide DNAm that included blood samples taken from nearly 16,000 participants included in the Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium (including 2,433 current, 6,518 former, and 6,956 never smokers) [509], Joehanes *et al*. (2016) identified 2,623 CpG sites, annotated to 1,405 genes, that were differentially methylated between current and never smokers, based on a Bonferroni-adjusted $p<1\times10^{-7}$. When a less stringent False Discovery Rate (FDR) of <0.05 was applied, over 18,700 CpGs were detected [509]. Of these sites, 185 were significant ($p<1\times10^{-7}$) in former versus never-smokers, providing evidence of both a persistent pattern of altered methylation in former smokers, and attenuation or reversal after smoking cessation. In addition, Joehanes and others have found that these methylation biomarkers are highly informative of several smoking-related traits, including pulmonary function, cancers, inflammatory diseases and heart disease [509 510].

Around the same time that Joehanes and colleagues published their findings, Lui *et al*. (2016) developed four separate DNAm-based alcohol models as biomarkers of alcohol intake using least absolute shrinkage (LASSO) regression techniques, also in the CHARGE consortium. The models comprised a set of 5, 28, 78 and 144 CpGs, respectively. Of these four models, the one corresponding to 144 CpGs provided the best discrimination, with AUCs upwards of 0.90 (in four replication cohorts) for current heavy drinkers versus non-drinkers, where the definition of 'heavy drinker' was men who drank ⩾42 g per day and women who drank ⩾28 g per day [511]. This analysis was well-powered and the proportion of variance in alcohol consumption explained by the DNAm biomarker was promising. However, Hattab *et al.* (2018) point out that the reported performance of the model may be inflated, since the authors did not use the coefficients from the discovery set, which was used to identify the 144 CpGs, to determine out-of-sample prediction [512]. Instead, the

authors ran their LASSO regression in their training dataset, which selected the variables that were important enough to remain in their model, then re-estimated the value of their coefficients by fitting separate logistic regression models in their replication cohorts.

In 2018, Dr Riccardo Marioni's group in the Centre for Genomic and Experimental Medicine described the developed of four DNAm-based predictors of complex traits using data derived from whole blood samples in two large cohort datasets, namely Generation Scotland (training dataset) and the Lothian Birth Cohort 1936 (test dataset). The traits of interest included alcohol consumption, smoking status, BMI and educational attainment. Again, using penalised regression techniques (i.e. LASSO), the authors identified predictors based on 287 (smoking), 371 (alcohol), 1,099 (BMI) and 281 (educational attainment) CpG sites. The DNAm-based predictors explained different proportions of the phenotypic variance, ranging from 2.6% for education to 60.6% for smoking. When the authors obtained AUC estimates for binary categorisations of these phenotypes, they found near-perfect discrimination between current and never smokers (AUC=0.98), whilst the discriminatory power for the identification of obese individuals (versus non-obese) and heavy drinkers (versus light-to-moderate drinkers) was moderate (AUCs of 0.67 and 0.74, respectively); poor discrimination was obtained for those with more years of full-time education (AUC=0.59). Interestingly, as with previous EWAS analysis of education [513], many of the CpGs identified overlapped with smoking-related DNAm signals, including cg11902777, which is located in the AHRR gene. In fact, this site had the fourth largest coefficient in the education DNAm LASSO model.

Overall, it appears that DNAm predictors are able to predict lifestyle factors with varying success. Smoking seems to be particularly amenable to detection using DNAm biomarkers, which may provide more accurate measurements than self-report, thereby improving disease prediction and risk stratification in both clinical and epidemiological settings. Moreover, unlike biochemical biomarkers such as cotinine, DNAm could serve as stable biomarker for lifetime exposure.

4.4.3. Epigenetic biomarkers of aging

As we age, the cells, tissues, and organs of our bodies undergo biological changes, which are accompanied by a progressive loss of function and an increased risk of morbidity and mortality. Chronological age is an imperfect surrogate measure of the aging process, however, because ageing does not affect people, or tissues, uniformly [514]. Instead, research suggests that these changes may be influenced by endogenous or exogenous stress factors [515]. The search for reliable indicators of biological age has been an active area

of research since the early 1980s [516]. In the last decade or so, a number of mathematical models predicting age from DNAm profiles, also referred to as “epigenetic clocks”, have been developed and this has largely been catalysed by the completion of the Human Genome Project, as well as a shift in scientific culture which has fostered the use of open access data sources [514]. A comparative review by Jylhävä *et al*. (2017), which considered six potential types of biological age predictors: epigenetic clocks, telomere length, transcriptomic predictors, proteomic predictors, metabolomics-based predictors, and composite biomarker predictors, concluded that predictors based on DNAm age (i.e. epigenetic age) currently perform best [516], although more longitudinal validation is required.

Epigenetic clocks are essentially a weighted measure of DNAm levels at specific CpGs. Selection of the most informative CpGs, along with their associated weights, is typically achieved using penalised regression models such as LASSO or elastic net, which, as touched on above, automatically screen for the best predictors [514]. Here, estimates are derived by regressing chronological age (dependent variable) on DNAm levels (covariates) [514]. Many of the CpGs that are incorporated into the algorithm provide limited predictive value on their own, however, the linear combination of CpG methylation beta values is highly correlated with chronological age. In addition, some clocks have been shown to predict aging-related diseases, such as coronary heart disease, diabetes, and some forms of cancer [517 518]. Different clocks exhibit varying degrees of accuracy depending on the set of CpGs used and the specific tissue.

4.4.3.1. Single-tissue DNA methylation-based age estimators

The first DNAm age estimator, described by Bocklandt *et al*. back in 2011, was built using DNA extracted from saliva (n=34 males) [519]. The model, which was based on methylation at just two CpG sites on the Illumina Infinium HumanMethylation27beadchip array (Illumina 27k array), explained 73% of the variance in age in an independent dataset (31 males and 29 females aged 18 to 70 years) and was able to predict an individual’s age with an average accuracy of 5.2 years. A couple of years later, Hannum *et al*. (2013) built a multivariable linear regression model for aging in whole blood samples (n= 656). “Hannum’s clock” integrated 71 CpG sites from the Illumina 450k array [520]. The model was accurate, with a correlation between age and predicted age of 96% in the training cohort and 91% in the validation cohort (n=171). The corresponding errors were 3.9 years and 4.9 years, respectively. Most of the methylation markers included in Hannum’s model are found in or near genes with known functions in aging-related conditions, such as Alzheimer’s disease and cancer. A limitation of Hannum’s clock is that it may be confounded by age-related

changes in blood composition [521] [522]; however, in blood, the model outperforms other 'multi-tissue' age-predictors [516].

4.4.3.2. Multi-tissue DNA-methylation based age estimators

Multi-tissues age estimators are so called because, unlike the models described above, they are applicable to all cell types and tissues across the entire lifecourse [514]. The obvious challenge when developing such a model is that there are well-established differences in DNAm patterns among different cell types and tissues, which may differ in early and later life. There are nonetheless age-related DNAm marks at specific sites across the genome, for instance within the bivalent chromatin domains and targets of Polycomb repressor 2 (*PRC2*), that appear to be conserved across different cell types [514]. The first, and arguably the most well-known multi-tissue age-estimator, was constructed by Horvath in 2013. "Horvath's clock' was trained on 7,844 samples, including 51 different (non-cancer) tissues and cell types, derived from 82 publicly available datasets (measured on the Illumina 27K or Illumina 450K array platform) [523]. To build the model, a transformed version of chronological age was regressed on the CpGs using elastic net regression, which selected a set of 353 CpGs. Using this transformed outcome measure, Horvath uncovered an interesting phenomenon: the rate of change of DNAm age, likened to the ticking rate of the epigenetic clock, slows down after adulthood. The high accuracy of this clock (age correlation 0.97 and 0.96, error = 2.9 years and 3.6 years in training and validation datasets, respectively) has been confirmed in hundreds of independent datasets [514].

4.4.3.3. Phenotypic age estimator

The so-called 'first generation' epigenetic clocks described above, i.e. the blood-based algorithm by Hannum and the multi-tissue algorithm by Horvath, were developed specifically to predict chronological age. Consequently, they only include CpG sites that exhibit strong time-dependent changes in DNAm. They do not capture CpG sites that account for differences in risk among individuals of the same chronological age [514] [524]. Therefore, whilst these DNAm-based age predictors are associated with many age-related diseases and outcomes, including life expectancy, the effect sizes are typically modest. Conversely, work by Levine and others has shown that "phenotypic aging measures", derived from clinical biomarkers, provide strong predictors of mortality risk and functional decline [524].

The phenotypic age estimator developed by Levine *et al* (2018) was also built using a penalised regression model, but instead of chronological age as the dependent variable, the authors constructed a surrogate measure of biological age based on a weighted average of

ten clinical characteristics: chronological age, albumin, creatinine, glucose, C-reactive protein levels, lymphocyte percentage, mean cell volume, red blood cell distribution width, alkaline phosphatase and white blood cell count [524]. This novel measure of 'phenotypic age', which was developed using clinical data from nearly 10,000 adults in the third National Health and Nutrition Examination Survey (NHANES III), signifies the age that corresponds with that person's mortality risk, based on the general population. So, a person with a phenotypic age of 55 years, for example, has the average mortality risk of someone who is 55 years old chronologically, irrespective of their own chronological age. These values were regressed on DNAm levels in blood from 456 participants in the Invecchiare Chianti (InCHIANTI) study, which automatically selected 513 CpGs (out of a possible 20,169 CpGs available on the 27k, 450k and EPIC arrays). The weighted average of the 513 CpGs yields a DNAm based estimator of phenotypic age, which the authors call '*DNAm PhenoAge'*.

DNAm *PhenoAge* outperforms Horvath and Hannum *DNAmAge* measures in predicting 10- and 20-year mortality and is strongly related to a number of age-related morbidity outcomes, including number of coexisting morbidities, likelihood of being disease-free and cardiovascular disease [524]. Moreover, DNAm *PhenoAge* has been associated with an increased risk of lung cancer incidence and mortality, after adjusting for chronological age, race/ethnicity, pack-years, and smoking status [524]. Similar to Hannum's clock however, DNAm *PhenoAge* was developed using whole blood and may therefore produce biased age estimates in non-blood tissue [514], though the authors provide empirical evidence to suggests that their age predictor performs remarkably well across a wide a range of different tissues and cell types [524].

4.4.3.4. GrimAge

In 2019, Horvath's group published a paper describing the development of a novel biomarker of biological aging that is a linear combination of chronological age, gender, and DNAm-based surrogate biomarkers for smoking pack-years and seven plasma protein levels [525]. The composite biomarker, named DNAm *GrimAge* owing to the fact that high values signify "grim news", outperforms existing DNAm-based biomarkers in terms of its ability to predict time-to-death.

DNAm *GrimAge* was developed using a two-stage approach. In the first stage, surrogate markers for pack-years of smoking, which is a significant risk factor for morbidity and mortality, and plasma protein levels were defined and validated using data from the Framingham Heart Study (n=2,356, split randomly into a training dataset, n=1,731, and a

test data set, n=625). In total, 12 of the 88 plasma protein levels (adrenomedullin, beta-2 microglobulin, cystatin C, growth differentiation factor 15, leptin, plasminogen activation inhibitor 1 and tissue inhibitor metalloproteinase 1) exhibited a correlation coefficient r >0.35 between their measured levels and their respective DNAm-based surrogate markers, when the analysis was restricted to CpGs that are present on the Illumina 450k array and the Illumina EPIC array. In the second stage, time-to-death (all-cause) was regressed on chronological age, sex and the above mentioned DNAm-based surrogates. The resulting values, i.e. the linear combination of covariates produced from the elastic net cox regression, were transformed to give units of years.

Consistent with previous publications, the authors defined an age-adjusted measure of *GrimAge* (i.e. *GrimAge* acceleration [*AgeAccelGrim*]), which, by definition is not correlated with chronological age. *AgeAccelGrim* is associated with a range of age-related conditions/outcomes, including but not restricted to, incident coronary heart disease, hypertension, lower physical functioning, type 2 diabetes and comorbidity index and time-to-cancer (any)

4.4.3.5. Epigenetic age acceleration

The discrepancy between DNAm age, as measured by the epigenetic clock, and chronological age may provide a reliable indicator of healthy aging [514]. When an individual's predicted methylation age exceeds their chronological age, implying that they are biologically older than their years, they are described as exhibiting "positive epigenetic age acceleration" [514]. The reverse situation, whereby an individual's tissue is aging slower than would be expected based on chronological age, would be described as "negative epigenetic age acceleration". Higher epigenetic age acceleration (EAA) has been associated with higher risk of all-cause mortality and poorer measures of physical and cognitive performance [517 521 522 526 527].

There are a several different measures of EAA described in the literature. Broadly speaking, EAA measures can be categorised into two groups: those that are independent of age-related changes in blood-cell composition, described in previous publications as 'intrinsic' epigenetic age acceleration (IEAA), and those that incorporate age-related changes in cell composition, referred to in previous studies as 'extrinsic' epigenetic age acceleration (EEAA) [514 522].

Briefly, IEAA is defined as the residual resulting from regressing epigenetic age on chronological age and measures of blood immune cell counts (naive CD8+ T cells, exhausted CD8+ T cells, plasmablasts, CD4+ T cells, natural killer cells, monocytes, and granulocytes). EEAA, by comparison, “up-weights” the contribution of three cell types whose levels are known to change with age (naive cytotoxic T-cells, exhausted cytotoxic T-cells), before regressing on chronological age [522] [528] [529]. As such, EEAA is able to capture aspects of immunosenescence, i.e. the age-related functional decline of the immune system.

Regardless of the type of measure used, age acceleration is useful because it identifies outliers or deviations from the norm, which may provide better prediction of age-related health outcomes than chronological age. The mechanisms that drive EAA are not well understood, however, Quach and others have shown that certain lifestyle factors can influence epigenetic aging rates [529]. Specifically, lower EEAA was found to be associated with lower BMI, higher education, higher fish intake, moderate alcohol consumption and higher blood carotenoid levels (an indicator of fruit and vegetable consumption) in data from older women within the Women's Health Initiative (WHI); lower IEAA was associated with poultry intake and lower BMI. Both EEAA and IEAA were found to relate to indicators of metabolic syndrome. The authors found no association with current smoking status.

A summary of the different EAA measured used in this thesis is provided in Chapter 9.

4.5. HPV detection methods in HNC

As highlighted in Chapter 3, the distinction between HPV- and non-HPV associated disease (i.e. cancers that are, and are not, linked with HPV infection) is important in relation to clinical prediction and disease management. Accurate detection of viral status is also essential for the purposes of research. As yet, no consensus has been reached on the optimal way to determine HPV status in HNC. The ideal test would be non-invasive, economical, easy to perform and interpret, and easy to incorporate into routine clinical practice. The objective of this section is to outline the main detection methods available and consider some of the advantages and disadvantages of each. An emphasis will be placed on HPV serological testing as this is the method used to determine HPV status in this thesis. Additional methodologies that are discussed include immunohistochemical (IHC) staining for p16, HPV polymerase chain reaction (PCR) testing and HPV in situ hybridization (ISH) analysis. An overview of this information, as well as information on some additional techniques that are not covered here, are presented in Table 17.

Before the different approaches are discussed, it is necessary to provide a brief description of the molecular pathogenesis of HPV infection. As mentioned previously, the viral proteins E6 and E7 are fundamental to the development of cancer. The E6 oncoprotein forms a complex with the cellular ubiquitin-protein ligase E6AP, which targets the TS protein p53 for ubiquitination and degradation [530]. This impairs normal cellular responses to DNA damage (i.e. G1 cell cycle arrest or induction of programmed cell death) and allows infected cells to proliferate. The E7 protein on the other hand, promotes transformation by binding to the retinoblastoma (Rb) protein, again targeting it for ubiquitination. Degradation of pRb releases E2F transcription factors from their negative control, which then stimulates the synthesis of enzymes needed to drive the cell forwards into S-phase [531 532]. The continuous degradation of pRb by HPV E7 also results in significant overexpression of the TS protein p16, via a feedback interaction [533].

4.5.1. p16 immunohistochemistry

Since HPV-driven carcinomas (i.e. carcinomas that express viral oncogenes) demonstrate overexpression of the p16 TS protein in response to loss of cell cycle control, p16 staining of tumour tissue by IHC presents a reasonable surrogate marker for HPV infection (14, 15). Within the UK, the National Institute for Health and Care Excellence (NICE) recommends p16 IHC for all cases of OPC [534]. Its main advantages are that it can be performed at low cost, has high sensitivity [94%], can easily be incorporated into clinical laboratories and can be performed on formalin-fixed paraffin embedded (FFPE) tissue blocks [534]. Some studies have also described the use of p16 IHC on fine needle aspirates (FNA), saliva, brush cytology and serum/plasma, thereby avoiding the need for surgical biopsy [535 536]. One potential limitation of the method is that interpretation of IHC staining can be subjective, however, only strong staining is considered indicative of HPV positivity (current UK guidelines recommend >70% staining). Examples of HPV-positive and HPV-negative tumours by IHC are depicted in Figure 20. The main limitations of p16 IHC are that it has lower specificity than alternative detection methods such as DNA ISH and DNA/RNA polymerase PCR, i.e. ICH analysis does not differentiate HPV serotypes, and p16 can be elevated by mechanisms other than HPV, leading to false positive results [537 538]. Indeed, some studies report that up to 20% of p16-positive OPCs are actually HPV-negative (20). On the other hand, the p16 (*CDKN2A*) TS gene is also one of the most frequently deleted genes in HNC and this could result in a false negative assessment of the presence of HPV [538].

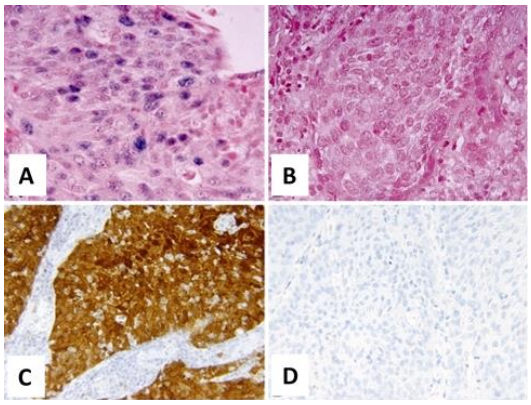
4.5.2. HPV in situ hybridization

DNA ISH, a method which is based on the use of radioactive or fluorescently-labelled nucleic acid probes that hybridize to target HPV DNA sequences, permits simultaneous identification and localisation of HPV DNA within a tissue sample [539]. It has the advantage of being highly specific in that probes can be designed to detect DNA sequences that are unique to individual HPV sub-types, in addition to sequences that are common to multiple subtypes. One of the major limitations of ISH is that it lacks sensitivity when HPV DNA copy numbers are low. The limit of its sensitivity is generally recognised as 10 viral copies per cell, however with the development of improved reagents and techniques, it is now possible to detect as few as one to two copies of HPV DNA per cell. In addition to potential sensitivity issues, ISH cannot provide direct evidence of transcriptional activity, even though it can differentiate between episomal and integrated DNA, since the presence of DNA does not necessarily indicate viral gene expression. Detection of mRNA E6/E7 transcripts is generally regarded as the "gold standard" for HPV detection, however the approach is technically challenging to perform and requires complex tissue processing (e.g. microdissection of fresh frozen tissue). For this reason, its use has traditionally been restricted to research laboratories [539 540]. Recent advances, including the development of RNAscope [540-542], permit visualisation of viral transcripts in routine clinical testing i.e. in FFPE samples [543], but so far the platform is not widely available in diagnostic laboratories.

4.5.3. HVP polymerase chain reaction testing

PCR is a type of targeted amplification technology which duplicates fragments of DNA, in this case HPV DNA, from a known sequence, thereby providing concentrated samples [544]. The inherent strength of this technique is that it requires small amounts of DNA and the technology is also widely available, making it a popular method of HPV detection in epidemiologic investigations. PCA has two main disadvantages. First, like DNA ISH, it cannot distinguish 'clinically significant' HPV infections from 'non-clinically significant' infections (i.e. HPV infections that results in transcriptional activity), although quantitative real-time PCR (Q-PCR) does permit assessment of HPV-16 viral load, which may suggest active replication. Second, samples are easily contaminated by previously amplified specimens or from surrounding non-neoplastic/stromal tissue that is HPV infected, leading to false-positives [539 543]. Regarding viral gene expression, the presence of E6 and E7 mRNA positivity can determined using reverse transcriptase PCR (RT-PCR), a technique which creates cDNA sequences from mRNA and subsequently amplifies them using traditional PCR, but this approach is time-consuming and traditionally requires fresh frozen (FF) tissue [539].

Figure 20: Microscope images of HPV-positive and HPV-negative oropharyngeal tumours by ISH and p16 IH.



***A:** HPV-positive tumour by ISH (positive staining identified as blue nuclear dots) **B:** HPV-negative tumour by ISH. **C:** p16-positive tumour by IHC, showing strong and diffuse nuclear and cytoplasmic staining in 70% or more of the tumour cells. **D:** p16-negative tumour by IHC staining. Credit: Chernock et al (2009) [545].*

4.5.4. HPV serological testing

Serum antibodies to HPV proteins, especially the E6 and E7 early proteins, have been detected in people with several HPV-associated cancers [546]. For OPC specifically, studies conducted in Europe and the US report that 35%-42% of people with OPC are HPV-16 E6 seropositive up to 10 years before diagnosis, compared to fewer than 1% of healthy controls (yielding ORs as high as 274; 95% CI: 110 to 681) [547 548], and 85% are seropositive at the time of diagnosis [549]. As such, HPV-16 E6 seropositivity presents a promising biomarker for OPC detection and screening. Moreover, several studies suggest that pre- and post-treatment HPV-16 E6 antibody levels may be associated with risk of OPC recurrence [546 547 550-552].

Conventional serologic methods include enzyme-linked immunosorbent assays (ELISA) which are well-validated, having been used for over 30 years to examine antibody responses to bacterial/viral infections [553]. A schematic illustration of the ELISA is presented in Figure 21. The main limitation of this technology is that each sample can only be analysed for antibodies against one genotype of HPV in each well, resulting in low through-put of sera and high serum consumption [553 554]. Waterboer and his colleagues at the German Cancer

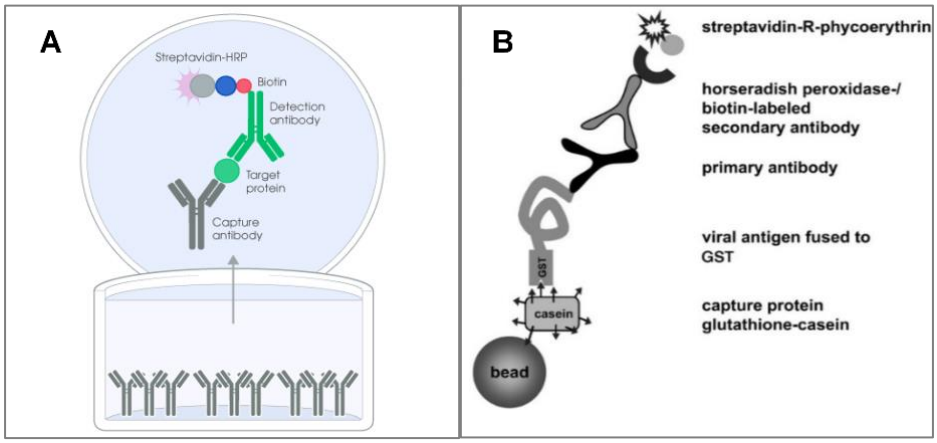
Research Center (DKFZ) have developed a high-throughput 'multiplex' HPV serology platform that allows simultaneous analysis of antibodies against multiple (>100) viral antigens in a single reaction [553] [554]. The assay, which uses Luminex fluorescent bead-based technology (Figure 21), can screen up to 1,000 sera per day.

This thesis utilises data derived from this novel multiplex serology platform. The protocol has been described in depth elsewhere [547] and further details will be provided in the next chapter; briefly, bead sets carrying different antigens (affinity-purified, bacterially expressed fusion proteins with N-terminal Glutathione S-transferase [GST]) are mixed with a small amount of serum in a single well-format and incubated; biotinylated secondary antibodies and fluorescent detection conjugates are then added to detect antibodies bound to beads and reporter fluorescence is determined using a flow cytometry-like analyser. Mean fluorescence intensity (MFI) values are dichotomized as antibody positive or negative using pre-defined cut-off values [555].

The multiplex serology assay is well-suited for epidemiological purposes because it is high-throughput, robust, requires small specimen volumes (50 µL), is relatively low-cost compared to alternative assays and can simultaneously measures responses to multiple HPV subtypes [554] [555]. One limitation, which is applicable to serologic tests in general, is that it is not site-specific and therefore it could be argued that infections outside the head and neck could influence the specificity of the assay.

In summary, there is no single test that has perfect specificity and sensitivity to detect HPV-associated tumours. The current 'research gold standard' test for HPV biologic activity, as suggested by the literature, is detection of E6/E7 viral transcripts using RT-PCR. However, this method is arduous to perform and the need for FF tissue is often cited as a reason why it is unsuitable for use in epidemiologic studies and clinical trials. For this reason, p16 IHC and HPV DNA ISH in combination are generally recommended. The recent development of a multiplex HPV serology assay provides an exciting addition to the field since HPV-16 E6/E7 antibodies may provide a tool for the early detection and/or prognosis of HPV-associated OPC. A commercially available assay is currently under development.

Figure 21: A comparison of antibody-capture enzyme-linked immunosorbent assay (left) and bead-based multiplex serology (right).



A: The basic set-up of a capture ELISA assay*. To detect viral proteins, a capture antibody, directed against the protein of interest, is first immobilised on a plate. The sample is added, and if viral antigens are present, they will bind to the immobilised antibody. The bound viral antigen is then detected using a biotinylated secondary antibody linked to a streptavidin-enzyme conjugated with alkaline phosphatase or horseradish peroxidase. A coloured product, proportion to the amount of the viral protein present in the sample is formed* [556] [557]*.*

B: A schematic representation of a bead-based multiplex serology assay. Binding of antigens is mediated by the interaction between the GST domain of the fusion proteins and glutathione on the glutathione-casein (GC) beads. Adapted from [558]*.*

Table 17: HPV detection methods in HNC.

Technique	Description	Advantages	Disadvantages
Routine histology	The presence of HPV is inferred from the tumour morphology. [534] [539].	• Universally available. • Low cost. • No additional equipment required.	• Small number of HPV-positive tumours do not exhibit typical features.
Southern blotting (SB)	Genomic DNA is extracted from a specimen and 'digested' by restriction enzymes. Fragments are separated by gel electrophoresis and transferred to a membrane before being hybridized with cloned HPV genomic probes [559].	• Well established. • Can detect as little as 0.1 copies of viral DNA per cell [559]. • Has the ability to differentiate between episomal and integrated DNA [559].	• Cannot be applied to FFPE samples. • Technically complex. • Requires a significant amount of DNA. • No practical/clinical utilization [559].
p16 IHC	Uses labelled antibodies to bind specifically to p16 antigens in situ.	• High sensitivity (94-100% %) [560] [561]. • High inter-observer agreement. • Relatively undemanding and inexpensive. • Readily available and interpretable (when signal intensity is high) • Has independent prognostic significance (in populations with high prevalence of HPV driven OPC).	• Surrogate marker for transcriptionally active HPV. • Lower specificity (79-82%) than ISH/PCR [560] [561]. • Subject to interpretation when staining is weak (<5% of cases) [562]. • Some studies show elevated expression of p16 in HPV DNA or E6/E7 transcript-negative OPCs [539]

Table 17 continued.

Technique	Description	Advantages	Disadvantages
DNA ISH	Uses labelled complementary DNA or synthetic oligonucleotides [i.e. probes] to localize to a specific HPV DNA sequence in a portion or section of tissue.	• High specificity [88-100%] [560] [561]. • Readily available probes for the detection of all known HPV strains. • FFPE tissue can be analysed for high-risk HPV sub-types using an automated platform and viewed with conventional light microscopy. • Can be used on FNA specimens. • 99% concordance between HPV detection in E6/E7 mRNA and HPV DNA ISH [540]. •	• Does not confirm transcriptionally active HPV. • Low sensitivity when DNA copy numbers in tumour tissue are low. • Low overall sensitivity (83-88%) compared to p16 IHC [560] [561]. • More expensive to perform compared to p16 IHC. • 11% reporter inter-observer variability [562].
RNA ISH	A labelled probe is used to hybridize to a known target RNA within a sample. The labelled probe is then detected using an antibody specific to the label on the probe.	• Confirms the presence of transcriptionally active HPV (superior to DNA ISH). • Amplifies low viral signal. • Lower risk of contamination (single test). • Easily interpretable results. • More sensitive and specific than DNA ISH (97 and 93% vs 88 and 88%).	• Technically difficult to perform. • Test is not routinely available. • Not available on an automated platform. • Limited clinical evidence of efficacy. • Not yet approved for clinical use.
HPV DNA PCR	A sequence of DNA is amplified using pre-specified HPV primers. The PCR products may be separated by electrophoresis or hybridised onto a chip.	• High sensitivity (97%) [560]. • Cost effective. • Numerous commercial assays available. • Can be used on FNA/saliva/brush cytology/serum/plasma specimens.	• Easily contaminated. • Detected DNA may not be from tumour tissue [559]. • May not correlate with actual HPV DNA cellular integration. • Provides no quantitative measure of viral load [559].

Table 17 continued.

Technique	Description	Advantages	Disadvantages
HPV mRNA PCR	Used to qualitatively detect gene expression through the creation of complementary DNA [cDNA] transcripts from E6/7 mRNA in combination with the amplification and detection steps of PCR.	• Provides evidence of transcriptionally active HPV. • High sensitivity.	• Expensive and time consuming. • Requires specialist research laboratory techniques. • As a consequence of RNA instability, testing has (until recently) relied on the analysis of fresh frozen tissue [561]. • Cannot localize HPV to the area of neoplasia. •
HPV-specific serum antibodies	Measures antibodies against early/late HPV capsid proteins.	• Does not require biopsy. • Individuals can produce a serologic response <10 years before.	• Currently no validated commercially available kits. • Serological biomarkers are not site-specific but can arise due to HPV infections at sites other than the head and neck (potentially affecting the specificity of the assay). • The significance of HPV seropositivity outside the oropharynx is unclear.

4.6. Metabolomic assessment of exposure

Chapter 3 summarised the published literature on the use of metabolomics as a tool to study HNC. The purpose of this section is to provide an overview of the different approaches used in metabolomics studies (i.e. target or untargeted) and describe the two most common analytical technologies for the generation of metabolomics data.

To recap, metabolomics is the study of the metabolite composition (including amino acids, organic acids, sugars, fatty acids, lipids, steroids, inflammatory markers, etc.) of a cell, tissue, or organism [563]. The total complement of low molecular weight compounds (metabolites) present within a biological sample is called the 'metabolome' and represents a "snapshot" of the response to of the biological system to both environmental exposures and upstream genetic, transcriptomic, and proteomic variation. As such, metabolomics offers a unique opportunity in which to examine the link between environmental and lifestyle exposures and health/disease outcomes. For example, previous metabolomic studies have identified both blood plasma and urinary metabolite correlates of traditional risk factors for cardiovascular disease, including blood pressure and hypertension [564].

Metabolomics experiments can be categorised as "targeted" or "untargeted", based on the study objective - discovery (untargeted) versus hypothesis testing (targeted) [563]. Targeted metabolite profiling techniques aim to quantify a predefined subset of metabolites, typically metabolites with related chemical structures and/or biological activities [563], whilst untargeted metabolomics studies aim to analyse of all the measurable metabolites within a biological sample, including unknown analytes. Untargeted metabolomic profiling studies often compare the metabolome of healthy and diseased groups (or control and test groups), in order to identify differences between them which may be relevant to specific biological conditions. By definition, untargeted metabolomics do not require any previous knowledge of the sample or a referential database. However, due to the comprehensive nature of untargeted metabolomics data, advanced chemometric techniques are needed to reduce it into a more computationally manageable set of signals [565 566]. The data analysis workflow for targeted and untargeted metabolomic studies have been described previously [565].

Two different but complementary technologies, which were introduced in Chapter 3, are recongnised as being particularly well-suited to analysing metabolites- NMR and MS [567]. NMR is based on the principle that when certain materials, such as tissue, are placed in a strong magnetic field, the atom nuclei take on a resonant characteristic, i.e. they can absorb

and then re-emit electromagnetic radiation at a specific frequency, and they become magnetized. The frequency of the radiation needed for absorption of energy depends on the chemical environment of the nucleus [568]. MS, by contrast, measures the mass-to-charge ratio of charged ions [569]. The choice of platform determines, to some extent, what can be measured, and each approach has its own advantages and disadvantages [567]. There are a number of excellent reviews published on the strength and limitations of NMR and MS as tools applicable to metabolomic analyses [570] [571]. Some of the main points are summarised in Table 18. In brief, NMR is a highly reproducible, non-destructive analytical technique that can detect many metabolites simultaneously in a short time period, but its sensitivity is low, and the apparatus is expensive to purchase and maintain. MS, by comparison, is inherently more sensitive with lower limits of detection, but sample preparation is more demanding, and samples cannot be recovered (i.e. they are destroyed), though only small amounts of sample are required [570] [571].

There are several methodological and statistical challenges when handling metabolomics data sets, which are pertinent to this thesis and which may affect the interpretation of results. They relate to the high-dimensional nature of the data, the biological variance and correlation structures of metabolite measures, not to mention analytical biases [572]. There are multiple methods available to help reduce complexity, identify variables of interest, and generate predictive models in multivariable datasets [572], some of which will be explored in later in this thesis.

Table 18: A comparison of the advantages and disadvantages of the two most common techniques used in metabolomics data acquisition.

Technology	Advantages	Disadvantages
MS	• High sensitivity- can detect metabolites in the femtomolar to attomolar range [567]. • Coupling MS with liquid chromatography (LC) or gas chromatography (GC) permits the measurement of hundreds of individual metabolite species within a single sample [567]. • Superior for targeted analysis [571].	• Moderate reproducibility. • More complex sample preparation required. • Precision - most studies rely on comparing peak area or intensity to locate differences in the relative abundance of specific metabolites between samples. • Quantification – signal intensity is affected by the type of sample preparation used and its molecular environment [567].
NMR	• High reproducibility. • Non-destructive – can recover the sample completely (and use it for MS). • Minimal sample preparation and no need for derivatisation. • Ease of automation. • Quantification - the peak area of a compound in the NMR spectrum is directly related to the concentration of specific nuclei [567]. • Versatility – can analyse metabolites in liquid state [serum, urine, plasma], in intact tissue (e.g. tumours), or *in vivo* [567].	• Low sensitivity- orders of magnitude less sensitive than MS [567] i.e. micromolar or millimolar concentrations. • Instrument is more expensive than MS.

4.7. Summary of the challenges and opportunities

One of the main limitations in epidemiological studies evaluating the association of exposures with outcomes is exposure misclassification. This may be especially true if the exposure or risk factor of interest is a health-related trait or behaviour, since these factors may differ across numerous dimensions (e.g. frequency or intensity of exposure) and occur over a prolonged period of time, as can be the case with smoking or alcohol use, for example. If these traits or behaviours are not measured accurately, this may result in a failure to detect an association between exposure and disease outcomes, or an attenuation of the relative risk. Of particular relevance to HNC, is how to accurately detect tumours that are HPV-driven. The validity of the exposure measurement in this case is improved by an understanding of the biology and etiologically of the agent. To this effect, serologic testing for HPV antibodies presents a promising research tool because the immune response to HPV16-driven tumorigenesis is detectable several years (i.e. over 10 years) before OPC diagnosis.

Self-report is a common data collection method in epidemiological studies. It is popular for several reasons; it provides a non-invasive and relatively cheap way of collecting information and, in many circumstances, has been shown to provide an accurate assessment of environmental and lifestyle exposures. Self-report is however subject to misclassification because it relies on accurate recall. Biological predictors, such as salivary cotinine as a biomarker for smoking, can provide a more accurate measure than self-report and they incorporate individual variability in metabolism. However, they may not be suited to studies examining the effects of lifetime exposure since they are often only present in the body for a short period of time.

It has become increasingly apparent that smoking and other environmental/lifestyle exposures, including aging, lead to genome-wide epigenetic alterations. A number of DNAm-based predictors of complex traits have been developed based on the findings of large EWAS studies. These novel DNAm-based biomarkers could be used in epidemiological studies to circumvent issues of recall-bias that limit self-report and help assess the totality of lifetime exposures. The proportion of the phenotypic variance in explained by these predictors is variable however, meaning that not all exposures may be amenable to detection using DNAm biomarkers. DNAm–based age, as predicted by the “Epigenetic Clock”, has been shown to predict chronological age with astonishing accuracy, though its greatest potential probably lies in its ability to provide biomarkers of accelerated aging, age-

related conditions, and longevity. It is important to remember however, that the amount of age-associated DNAm changes are specific to cell type and tissue and this should be factored into any study design. Overall however, epigenetic biomarkers may help address many long-standing questions in the field of HNC, including what the effect of established risk factors on survival is.

Finally, metabolomics, which measures the complement of small molecule-weight metabolites present in a biofluid or tissue, offers great promise for uncovering the causal relationships between environmental or lifestyle factors and disease outcomes. The complex and dynamic nature of the metabolome and the relative infancy of the field, however, present a number of statistical and analytical challenges.

This thesis will draw upon a combination of self-reported phenotypes, DNAm-based predictors of exposure, and NMR-derived metabolomics data to compare and quantify the effect of specific lifestyle traits and behaviours on HNC survival. HPV status will be determined using HPV serology assays. Further details of this and of the DNAm profiling techniques used in this analysis are provided in the next chapter.

Chapter 5: The Head and Neck 5000 study

The study population for this thesis was comprised of individuals enrolled in the H&N5000 study. H&N5000 is a large, UK-based, prospective observational study of people with HNC. This chapter provides a brief description of the participants included in the study and an overview of the data collection processes relevant to this thesis, including information on questionnaire design, DNAm and metabolic profiling. For a more comprehensive description of the study, including study design and recruitment, please see Ness *et al*. [573] [574]. Copies of the H&N5000 study documents, such as the consent form, data capture form and participant questionnaires, are available through the study website [315].

5.1. Study design and follow-up

All potentially eligible individuals were identified by the multi-disciplinary team (MDT) treating them. The inclusion and exclusion criteria for entry into the study are outlined below.

5.1.1. Inclusion criteria:

- Anyone with a new primary HNC (ICD-10: C00-C14, C32 and C73);
- People with an unknown primary cancer were eligible for inclusion if the MDT felt that the primary was likely to be HNC;
- Individuals aged 16 years old or over;
- Those who were already participating in other studies were still eligible for inclusion.

5.1.2. Exclusion criteria:

- Individuals who did not have HNC;
- Anyone with lymphoma, skin cancer, a secondary HNC or a recurrence of a previous HNC;
- Anyone who was considered to meet the criteria for mental incapacity or vulnerability;
- Anyone who had already commenced their cancer treatment, unless the treatment was their diagnostic procedure (e.g. tonsillectomy or thyroidectomy) or people were being treated palliatively, in which case individuals were recruited as soon after their diagnosis as possible (i.e. within one month of the procedure).

5.1.3. Consent

At the point of enrolment, i.e. before individuals started their cancer treatment, research nurses (or dedicated health professionals) at each site obtained written informed consent. The consent form asked participants to confirm that they were happy to complete questionnaires and that they gave their permission for the research team to: collect, store and use biological samples (blood and saliva); obtain samples of stored tissue; carry out genetic analyses; and collect information from hospital notes and via linkage to health-related records including disease registries [574].

5.1.4. Recruitment rates

Between April 2011 and December 2014, 5511 people with HNC, from 11,158 people who were identified as potentially eligible (49%), were recruited from 76 centres across England, Scotland and Wales [573] (Figure 22). There was variation in recruitment and response rates by H&N5000 study centres, with recruitment rates ranging from around 20% to around 90% of eligible HNC cases [573]. It is estimated that when all study centres were open, the study captured a third of all incident cases in the UK [575].

Of the 5511 participants recruited, 5,373 were confirmed eligible [576]. The reasons for ineligibility were:

- The tumour was not HNC (n=68)
- There was missing tumour site data (n=37)
- Clinical staging was 0 (n=18)
- The participant withdrew (n=63)

5.1.5. Baseline data collection

The various data collection points are illustrated in Figure 23. Once informed consent had been obtained, participants were given three questionnaires to take away and return by stamped address envelop or complete in clinic. The first included questions on socio-economic position (including occupation, education and housing) and lifestyle (including tobacco use and alcohol intake); the second included questions on sexual history (e.g. timing of first sexual intercourse, number of sexual partners), and the third asked about psychological status and QoL (general as well as cancer-specific).

During the same clinic appointment, participants were asked to provide a blood and saliva sample, which were collected using a standardised protocol. Further details are provided

below. If an individual did not wish to provide a sample, they remained in the study for questionnaire completion and data capture. All baseline questionnaires and blood and saliva samples were done before the participant started their HNC treatment, with the exception of those whose treatment was their diagnostic procedure and those who were being treated palliatively. Where possible, paraffin embedded tissue blocks were obtained from local pathologists, along with an anonymised copy of the participants' histopathological report. Tissue samples were not used in the current analysis, but further details are available on the study website.

Data capture forms (DCFs) were completed by research nurses at each participating site, using information extracted from participants' medical records (for those participants who had consented). The data extracted included information on diagnosis (e.g. date of diagnosis, ICD code, histology), treatment (e.g. cancer care plan intent and sequence) and existing comorbidity (which did not include the index cancer).

The intended cancer care plan for the patient's HNC at baseline could be either curative, palliative, supportive, The terms "palliative" and "supportive" are often used interchangeably and both are given in the knowledge that the cancer cannot be cured; the distinction is that supportive care refers explicitly to all non-cancer focused treatment i.e. it is just intended to prevent and/or relieve the symptoms of cancer (e.g. hospice care), whilst palliative anti-cancer care is given in combination with anticancer treatment (e.g. chemotherapy or radiotherapy) with the aim of extending an individual's life for as long as is possible and comfortable. If a participant decided that they did not want to receive any treatment, this would be entered in the DCF as "No specific anti-cancer plan".

Comorbidity was defined using the Adult Comorbidity Evaluation-27 (ACE-27) index [577], which was described in Chapter 3. ACE-27 is a 27-item assessment tool that has been validated for use in multiple patient groups [396 578 579]. Using the tool, nurses graded participant's comorbidities into one of four categories according to the severity or organ decompensation: none (coded 0), mild (coded 1), moderate (coded 2), or severe (coded 3). An overall comorbidity score was assigned according to the severity of the highest ranked medical condition, excepted in cases with two or more grade 2 ailments in different organ systems, in which instance a final score of three would be assigned. The comorbidity grade that corresponded to final comorbidity score was entered into the DCF (i.e. no-comorbidity, mild decompensation, moderate decompensation, or severe decompensation, or 'unknown'). Research nurses were not asked to record the underlying conditions that resulted in this score, and as a result, responses could not be cross-checked or grouped in any way.

5.1.6. Response rates

Response rates to the baseline health and lifestyle questionnaire ranged from <30% to >90% and the percentage of participants providing a blood sample varied from <50% to >90% [573]. Altogether, 5,474 (99%) DCFs and 4,099 (74%) health and lifestyle questionnaires were completed [573]. As of May 2019, there were 3,391 people with valid stage, baseline health and lifestyle data and a blood sample available [576]. There are 2,993 participants with tissue blocks (most are primary HNC, but some are neck node; personal communication).

5.1.7. Follow-up

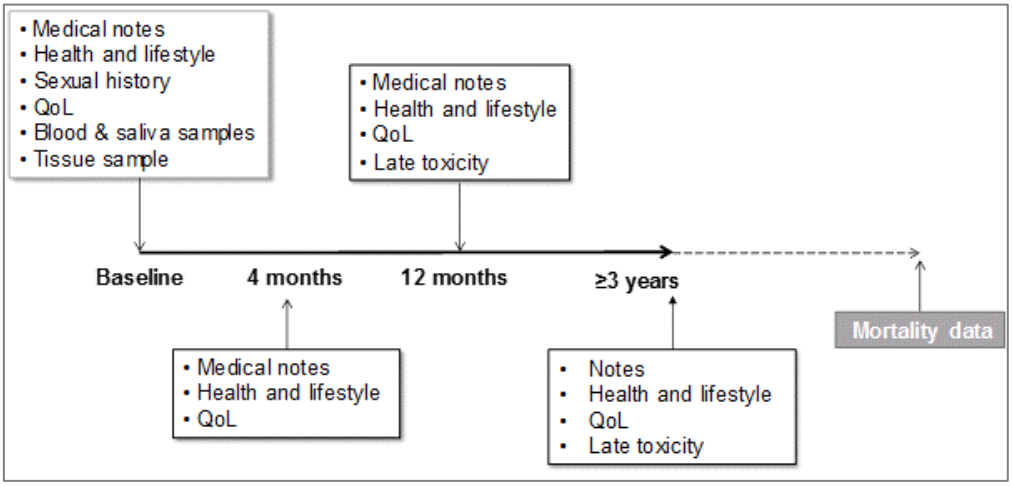
At four months and twelve months after enrolment, participants were sent a follow-up questionnaire pack. The follow-up questionnaire included questions around concerns (e.g. fear of cancer recurrence), loss of function (e.g. as speech or swallowing), treatment received, personal costs (e.g. time off work, travel expenses or home help) and QoL. Further medical data were extracted from the hospital notes by research nurses. Between November 2016 and April 2019, participants who had been in the study for a minimum of three years were asked to complete an extended lifestyle questionnaire that contained additional questions around smoking and alcohol drinking histories (including behavioural change), marijuana use, tonsillectomy and dental health, amongst other things. Of the 76 H&N500 original centres, 63 agreed to take part in the extended follow-up. To date (May 2019) 2,185 of 3,540 (62%) of questionnaires and 4,328 (91%) of DCF have been completed and returned to the study team. Information obtained from these questionnaires is not used in the current analysis but will form the basis of future research.

Throughout the study, the H&N5000 study team obtain on-going updates from the NHS Central Register (NHSCR) and NHS digital (formerly called the Health and Social care information Centre [HSCIC]) informing them of any new cancer diagnoses or deaths among participating individuals. Specifically, the study team receive information on date of death, place of death and cause of death, as recorded when deaths are certified and registered.

Figure 22: The 76 H&N5000 Study centres [580].

Image source: [580]

Figure 23: H&N5000 data collection points.



*Abbreviations: **QoL**, quality of life.*

5.2. Methods for collecting and processing materials

5.2.1. Blood sample collection, processing, and storage

At baseline outpatient clinics, research nurses or trained phlebotomists collected 16 ml of venous blood from participants, in accordance with the local standard operating procedures for drawing blood. Sample were collected in two EDTA tubes (10 and 6 ml each) and labelled with the participant's study ID number and a numeric ID barcode label. All samples were posted first class to the Bristol Bioresource Laboratories (https://www.bristol.ac.uk/population-health-sciences/research/groups/bblabs/] at ambient temperature, using the transfer kits provide by the H&N5000 study team. Over 60% of samples arrived within 48 hours and over 85% within 72 hours. On receipt, samples were checked for any sign of damage. Laboratory staff ensured that both EDTA tubes were labelled and that the accompanying transfer sheet displayed the same sample ID. Blood samples were separated using a Thermo Scientific Heraeus Megafuge 16 centrifuge and spun at 3500rpm for 10 mins at 4-5^{0}C. Up to 10 x 500 µl and 15x 200µl were aliquoted from the 10ml and 6ml EDTA tubes and placed in 0.5ml apex tubes. Where insufficient sample was available to obtain all aliquots, a mixture of 500µl and 200µl aliquots were taken. Plasma samples were kept for biochemical, proteomic and metabolomic measures. The buffy coat layer (the fraction that contains most of the white blood cells and platelets) was extracted from each EDTA tube and put into separate 2ml sterile tubes for DNA extraction. All samples were stored at -80 °C. The freezers in the repository are alarmed and covered by a 24hr a day call out system. In total, 4,587 baseline blood samples were obtained (oral cavity, n=1,147; oropharynx, n=1,611; larynx, n=924).

5.2.2. Genotyping and imputation

Samples were genotyped using the Illumina "OncoArray", which was specifically designed by the OncoArray Consortium (part of the Genetic Associations and Mechanisms in Oncology (GAME-ON) consortium) to evaluate genetic variants associated with common cancers [581]. The custom array, which covers approximately 600k SNPs, comprises a genome-wide backbone, which provides coverage of most common genetic variants, together with specific markers of interest for each of the five GAME-ON cancers (breast, colon, lung, ovary, and prostate), and SNPs associated with multiple common cancer phenotypes and risk factors (e.g. radiation response and BMI).

Genotype calls were made by the INHANCE Dartmouth team using GenomeStudio software (Illumina, Inc.). Initial quality control (QC) steps and analyses were performed at IARC, Lyon. After removing duplicates, related samples, samples with sex discrepancy and population outliers, imputation of unknown genetic variation was performed using the Michigan Imputation Server [582]. Genotypes were pre-phased (i.e. their haplotypes were inferred) using SHAPEIT v2 [583] and imputed with Minimach v3 [584] using the Haplotype Reference Panel (HRP) [585], which is a large collection of human haplotypes (n = 64,976) obtained from multiple initiatives including the 1000 genomes project. After imputation, SNPs with an imputation quality (R_2) lower than 0.7 were removed from the datasets.

5.2.3. DNA methylation profile generation

Genome-wide methylation status was assessed in a subgroup of participants with oropharyngeal tumours. Individuals were selected on the basis that they had OncoChip genotype data (see above), baseline questionnaire data and data capture information available (i.e. baseline questionnaire and DCFs had been completed and returned to the H&N5000 study team). Determination of tumour site was based on clinical ICD codes. Following extraction, genomic DNA (isolated from buffy coats) was first bisulphite-treated using the EZ DNA Methylation™ kit (Zymo, Irvine, CA, USA), which converts unmethylated cytosine into uracil, leaving 5-methylcytosine (5-mC) unchanged. Genome-wide methylation status was then measured at >850,000 cytosine positions across the genome using the Infinium MethylationEPIC BeadChip [586]. Full details of the array have been published elsewhere [587]. Briefly, the bead chip, which can run multiple samples simultaneously, contains a series of microwells and inside these wells are oligonucleotide probe sequences attached to silicon beads. The bisulfite converted DNA was first hybridised to the array, then single-base extension of the probe was used to incorporate fluorescently labelled dideoxynucleotides triphosphates (ddNTPs) at the 3' CpG. The bead chip was scanned using an Illumina iScan, which uses a laser to excite the fluorophore of the single-base extension product on the beads. The iScan output includes high-resolution images of the light emitted from the fluorophores. Initial quality review of the intensity signals was assessed using GenomeStudio. The raw data (IDAT files) from GenomeStudio were then imported into R and pre-processed (i.e. quality control and normalisation steps were performed) using the meffil package (available at https://github.com/perishky/meffil/). Firstly, for every sample, a detection *p*-value was generated for each CpG, which provides an indication of the quality of the signal. The detection *p*-value compares the total signal ($M + U$, where M and U refer to the average fluorescence intensity from the methylated and unmethylated target CpG respectively), to the background signal level, which is estimated

from the negative control probes. Very small *p*-values suggest a reliable signal whilst large *p*-values (>0.1), usually indicate a poor-quality signal. Mean detection *p*-values were plotted for each sample and those with a high value, implying many failed probes, were excluded from further analysis. Raw probe intensities were normalized (to reduce technical variability and batch effects) using functional normalization in order to minimise the non-biological differences between probes, as described in Min (2018) [588].

The proportion of DNA methylation present at each CpG site was reported as a β-value, which is ascertained by taking the ratio of the methylated probe intensity and the overall intensity: $\beta = M / (M + U + \alpha)$, where $\alpha = 100$ (to protect against division by zero). The β-value can range between 0 and 1, with 0 indicating a completely unmethylated CpG site and 1 indicating a fully methylated site [589]. At the time of writing this thesis, DNA samples have been analysed for 448 participants. Of these, 440 passed quality control (2 samples with incorrect sex prediction, 3 samples with sex detection outliers, 1 sample with an outlier in predicted median methylated vs unmethylated signal, 2 duplicate samples). An additional 32 samples were subsequently excluded owing to pathological re-classification, leaving 408 participants with epigenetic data available for analysis.

5.2.4. Metabolite quantification

Metabolic profiling was done on all participants with oropharyngeal tumours who had adequate blood available for analysis. Aliquots of stored plasma samples were analysed using an automated high-throughput NMR spectroscopy-based platform (Nightingale Health Ltd, Helsinki, Finland), which is hosted by the Department of Chemistry, University of Bristol. Details of the NMR platform have been published previously [590 591]. Briefly, 70 µL of plasma and 70 µL of sodium phosphate buffer (75 mM Na2HPO4, 0.08% sodium 3-(trimethylsilyl) propionate-2,2,3,3-d4, 0.04% sodium azide in 80%/20% H20/D2O, pH 7.4) were first mixed together using an automated Gilson 215 Liquid Handler. The resulting solution was transferred to 96-format racks of NMR tubes using a Varispan Janus liquid handling robot (PerkinElmer). The sample racks were inserted into one of the five well-plate positions in the SampleJet™ (Bruker BioSpin GmbH, Germany) sample changer, which sits on top of the superconducting magnet, where the actual NMR measurements take place. To prevent degradation of samples whilst they await measurement, the sample charger includes a cooling unit that keeps the prepared samples at refrigerator temperature (6 °C). The NMR spectra were acquired using a Bruker Avance III HD 600MHz spectrometer, equipped with a cryogenically cooled triple resonance probe head (CryoProbe Prodigy TCI). The profiling approach in based on three ‘molecular windows’, two of which (LIPO (lipoproteins) and

LMWM (low-molecular-weight molecules) were applied to the native plasma and one to the plasma lipid extracts (LIPID). NMR spectra are automatically transferred to a centralised server, which performs a number of additional spectral processing steps, including an overall signal check for missing or extra peaks, background control, baseline removal, and spectral area-specific signal alignments. In addition to this, spectral information is also compared against the spectra of the two quality control samples. Regression modelling is then performed to produce the quantified molecular data, as described previously (11,12,16). The combination of the three molecular windows captured by the platform provides simultaneous quantification of over 200 metabolic measures for each sample, including 14 lipoprotein subclasses, multiple fatty acids, glucose, various glycolysis related measures, ketone bodies, and amino acids in absolute concentration units [590-592]. The 14 lipoprotein subclasses include [593]:

- six sub-classes of very low-density lipoproteins (LDLs):
  - "extremely large", with an average particle diameter of ≥75 nm,
  - "very large" (64.0 nm),
  - "large" (53.6 nm),
  - "medium" (44.5 nm),
  - "small" (36.8 nm),
  - and "very small" (31.3 nm);
- three subclasses of low-density lipoproteins (LDLs):
  - "large" (25.5 nm),
  - "medium" (23.0 nm),
  - and "small" (18.7 nm);
- intermediate density lipoproteins (IDLs), with an average particle diameter of 28.6 nm;
- and four subclasses of high-density lipoproteins (HDLs):
  - "very large" (14.3 nm),
  - "large" (12.1 nm),
  - medium" (10.9 nm),
  - and "small" (8.7 nm).

For each of the 14 lipoprotein subclasses, the following measures are provided:

- the circulating concentration of total lipids in the particles (sum of free and esterified cholesterol, triglycerides and phospholipids),
- the particle concentration,

- the absolute circulating concentration of five main lipids (free, esterified and total cholesterol, triglycerides and phospholipids),
- and the relative proportions of these five lipids in each particle subclass [594].

This NMR metabolomics platform has been widely applied in genetic and observational epidemiological studies [593-603]. Indeed, as of March 2018, there were over 100 scientific articles published in biomedical journals that had applied this technology [604]. One of the main advantages of this approach is that in comparison with standard clinical chemistry assays, the NMR metabolomics platform can facilitate simultaneous quantification of many more biomarkers in a single experiment, whilst still preserving the accuracy achieved using routine clinical assays[594]. This makes it a very cost-effective method for the identification and validation of biomarkers in large-scale epidemiologic studies. In addition to this, NMR is highly reproducible and because samples never come into contact with the radiofrequency detector in the NMR spectrometer, the technique suffers no discernible batch effects, which can be an issue in studies that use alternative MS approaches (15). A more comprehensive comparison of the two approaches (NMR and MS) was provided in Chapter 3.

5.2.5. Determination of HPV status

As noted in the previous chapter, there are several different HPV detection methods available. The gold standard test is the detection of viral DNA and viral RNA in tumour tissue because this provides evidence that the viral DNA is incorporated and actively transcribed (i.e. tumour is HPV driven). However, the choice for an HPV test should be driven by practical considerations and by the intention for its use. For example, HPV serology has been shown to provide a robust surrogate marker for HPV driven cancer in the absence of suitable tumour tissue [548 605].

Various HPV detection methods were used in H&N5000. This thesis focuses on HPV serological data specifically, but further molecular markers including HPV DNA and RNA and cellular protein $p16_{ink4a}$ were measured on <1000 formalin fixed tissue blocks; blocks were selected on the basis of their HPV serology and location. For the serological testing, baseline blood samples were sent on dry ice to the German Cancer Research Centre (DKFZ) in Heidelberg and analysed using multiplex assays [606], a bead-based high-throughput hybridization technique that facilitates the simultaneous detection and genotyping of multiple HPV types. The method has been described in detail elsewhere [547]. Briefly, plasma samples were analysed for antibodies to the major capsid protein (L1), the early oncoproteins (E6, E7), and other early proteins (E1, E2, E4) of the following carcinogenic HPV subtypes: HPV16 and HPV18 (L1, E1, E2, E4, E6, and E7); HPV31, HPV33, HPV45,

and HPV52 (L1, E6, and E7). Further details are provided in the methods section of this thesis.

5.3. Variable description

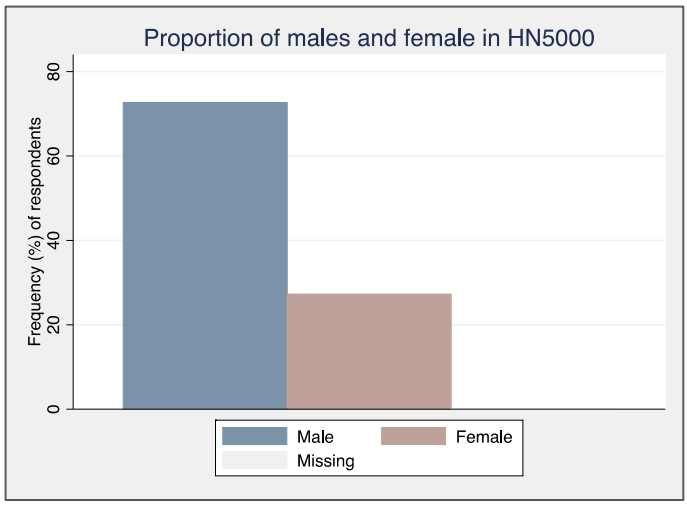
The following section of this chapter provides a description of the 'core' phenotypic variables used in this thesis. Core variables refer to those variables that are included in all H&N5000 collaborator datasets; they capture data collected from baseline questionnaires and DCFs. Additional variables have been derived for the purpose of specific analysis. Graphical summaries of each of the variables are provided to highlight the distribution (continuous variables) and proportion (categorical variables) of values within each category. These figures are based on the entire H&N5000 cohort (data release v2.5). A summary of the baseline descriptives of participants included in each of the analytic datasets is provided in the next chapter.

5.3.1. Demographic variables

5.3.1.1. Gender

Gender is a binary variable that takes a value of 1 for males and 2 for females. It is based on respondents' own self-reported gender, as reported in question *A4c* of the baseline *About You* questionnaire. Participants were predominantly (72%; n=3,928/5,402) male (Figure 24).

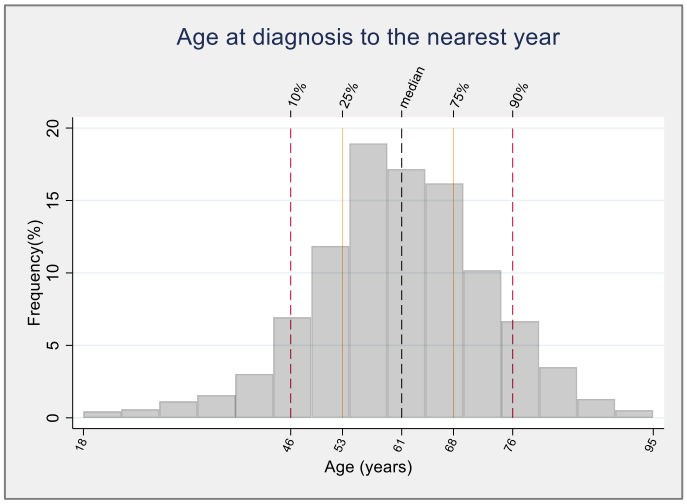
Figure 24: Proportion of male and female participants in H&N5000.



5.3.1.2. Age at consent

Participant's age is captured in a continuous variable. It was derived from question *A2* of the baseline *About You* questionnaire, which asks respondents to provide their date of birth using the format: day/month/year. The median age of participants was 61 years (interquartile range (IQR) = 53 to 68), and the range was 18–95 years (Figure 25)

Figure 25: Histogram showing age distribution in H&N5000.



5.3.1.3. Ethnicity

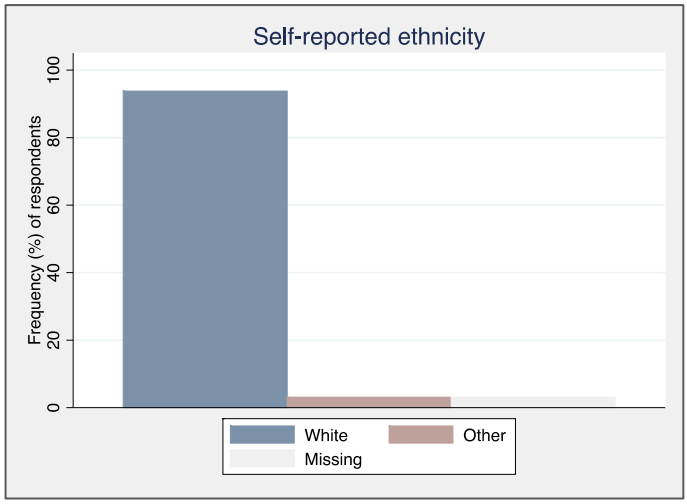
Participants' ethnicity was obtained from their medical notes and entered into the DCF (section A9). Research nurses were asked to enter an ethnicity code based on one of the following groups:

1) White –British
2) White -Irish
3) Any other White background
4) Mixed -White and Black Caribbean
5) Mixed -White and Black African
6) Mixed -White and Asian
7) Any other Mixed background
8) Asian -Indian or British Indian
9) Asian -Pakistani or British Pakistani
10) Any other Asian background
11) Black -Caribbean or British Caribbean
12) Black -African or British African

13) Any other Black background
14) Chinese
15) Any other Ethnic group
16) Not stated/given
17) Patient refused

Approximately 97% (n=5,068/5,235) of participants were white (Figure 26).

Figure 26: Bar chart of ethnicity in HN5000.



5.3.2. Clinical variables

5.3.2.1. Primary Diagnosis

A grouped ICD variable was created based on the ICD-10 classification codes described in Chapter 2:

Oral cavity	C00, C00.1, C00.3, C00.4, C00.9, C02.0, C02.1, C02.2, C02.3, C02.8, C02.9, C03.0, C03.1, C04.0, C04.1, C04.8, C04.9, C05.0, C05.2, C05.9, C06.0, C06.1, C06.2, C06.8
Oropharynx	C01, C02.4, C05.1, C05.8, C09.-, C09.0, C09.1, C09.8, C09.9, C10.-, C10.2, C10.3, C10.8, C10.9
Nasopharynx	C11.0, C11.1, C11.2, C11.3, C11.8, C11.9
Hypopharynx	C10.0, C12, C13, C13.0, C13.1, C13.2, C13.8, C13.9
Larynx	C10.1, C32, C32.0, C32.1, C32.2, C32.3, C32.8, C32.9
Thyroid	C73

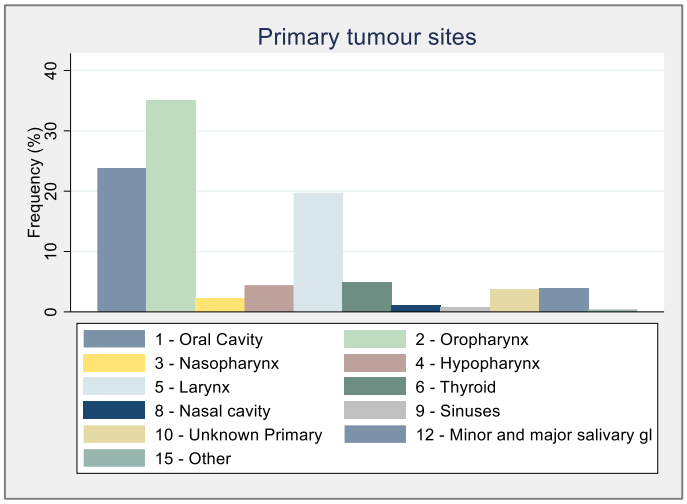
Salivary gland	C06.9, C07, C08.-, C08.0, C08.1, C08.9
Nasal cavity	C30.0
Sinuses	C31.-, C31.0, C31.1, C31.2, C31.3, C31.8, C31.9
Unknown primary	C76, C76.0, C80.-, C80.0, C80.9

N.B. these definitions are slightly different to those described by Conway et al (see Chapter 2), which were published later. Using the newer classification, pharynx, NOS (C14.0) and Waldeyer's ring (C14.2) are grouped under oropharynx wilst soft palate NOS (C05.1) and overlapping lesion of the palate (C05.8) are classified as oral cavity [15].

In the first H&N5000 data release, ICD codes were determined using the information provided in the baseline DCF, which was based on clinical examination/histology (i.e. pre-treatment). In the second data release, which was used in the current analyses, an updated ICD variable was made available. The updated variable used pathologically confirmed ICD codes in the first instance; where pathological reports were unavailable (n=763), clinical ICD codes were used. For 290 people, the pathology ICD code did not concur with the clinical ICD code. In such situations, individual cases were reviewed by Dr Miranda Pring (Consultant Senior Lecturer in Oral Maxillofacial Pathology) and Professor Steve Thomas (Professor in Oral and Maxillofacial Surgery) and assigned a "best fit" ICD code.

As illustrated in Figure 27, the most common tumour sites, using the revised ICD codes, were oropharynx (35%; n=1,896/5,404), oral cavity (23%; n=1,288) and larynx (20%; n=1,065).

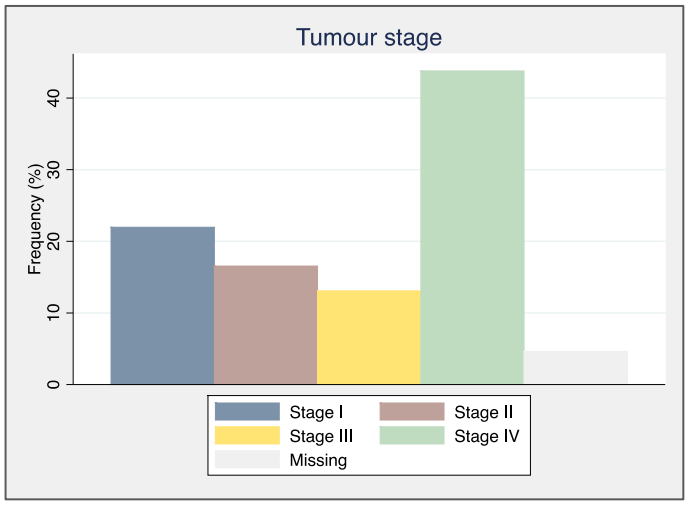
Figure 27: Proportion of HNCs by sub-site in H&N5000.



5.3.2.2. TNM staging

Based on the initial clinical ICD codes, a simplified staging variable was created which collapsed the eight possible stage groups outlined in Chapter 1 into four groups: I, II, III and IV, i.e. all stage 2 tumour groups were combined into one category and all stage 4 tumour groups were combined into one category. Separate codes were entered for incomplete T and/or N and/or M codes or unacceptable TNM combinations. When the ICD coding variable was updated using pathologically confirmed ICD codes (i.e. in the second data release), the grouped TNM variable was also updated to provide the 'best' (i.e. post-treatment) staging information available. Most cases were diagnosed at a high tumour stage (stages III and IV), with almost half (46%; n=2,366/5,154) being diagnosed at stage IV (Figure 28)

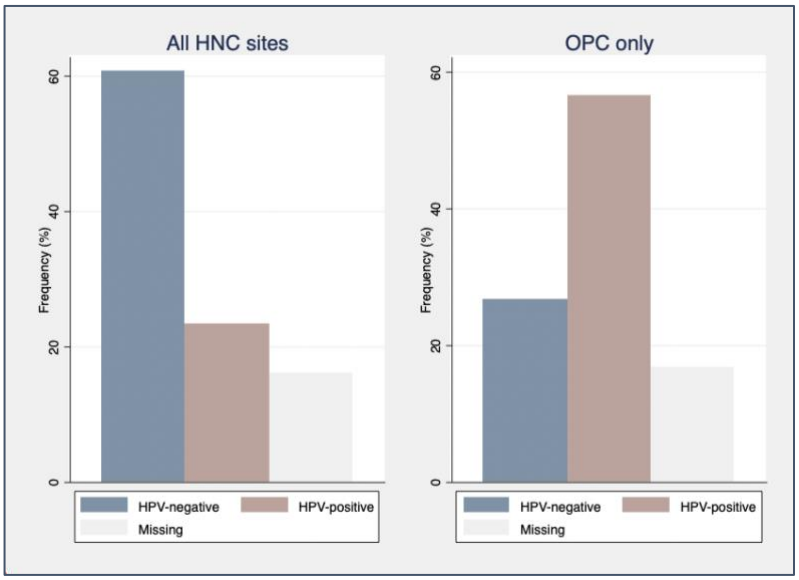
Figure 28: Bar chart showing the proportion of HNCs in each TNM stage group in H&N5000.



5.3.2.3. HPV status

MFI values were dichotomized as antibody positive or negative using predefined cut-offs. HPV16 E6 positivity was defined using a cut-off of > 1000 MFI. A binary variable was created to capture HPV status (positive or negative) based on this threshold. In total for all oral cavity, oropharyngeal and laryngeal cancers combined, 72% (n=3,274/4,541) of cases were HPV-negative (Figure 29). When restricted to oropharyngeal tumours specifically, 67.97% (n= 1,074/ 1,580) of participants were HPV-positive (Figure 29).

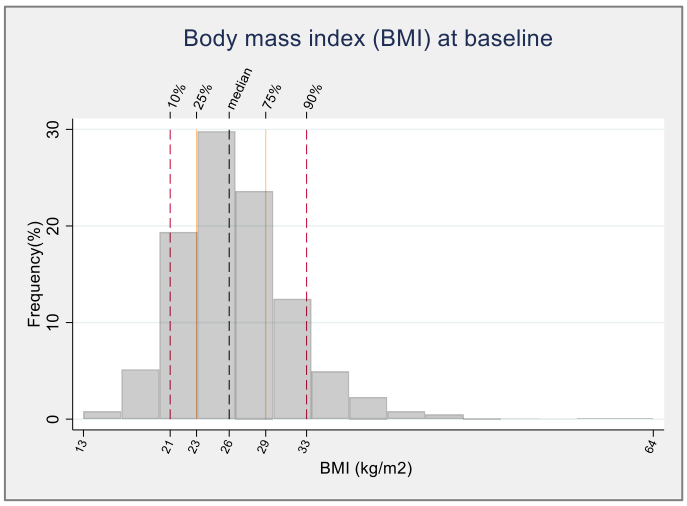
Figure 29: Bar chart showing the proportion of HN5000 participants who were HPV-positive or HPV negative, as determined by HPV-16 seropositivity.



5.3.2.4. Body mass index

BMI is a measure of weight relative to height and is defined as the body mass divided by the square of the body height (weight (kg) / (height (m))$_2$). The values needed to derive this measure were obtained from questions *A3* and *A4* of the *About You* questionnaire. Participants' could provide their height in centimetres (cm) or feet and inches and their current weight in kilograms (kg) or stone and pounds (lbs). Using the formula defined above, his information was used to create a continuous variable for BMI at time of diagnosis. Early versions of the baseline questionnaires did not enquire about participants' height and weight and as a result, BMI data is missing for just over 40% of participants. The median BMI for H&N5000 participants was 26 (IQR=23 to 29). The range was 13 to 64 (Figure 30)

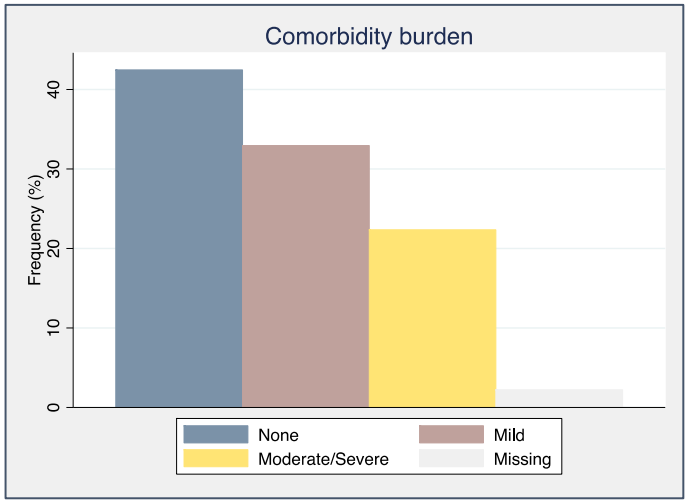
Figure 30: Histogram of BMI distribution in H&N5000.



5.3.2.5. Comorbidity

Based on the ACE-27 scores, which were described earlier in this chapter, a categorical comorbidity variable was derived corresponding to extent of decompensation- “none”, “mild”, “moderate or severe”. Around 43% (n=2,295/5,284) of participants had no comorbid illness, other than their HNC (Figure 10); 34% (n=1,781) had mild comorbidity and 23% (n=1,208) had moderate or severe comorbidity (Figure 31).

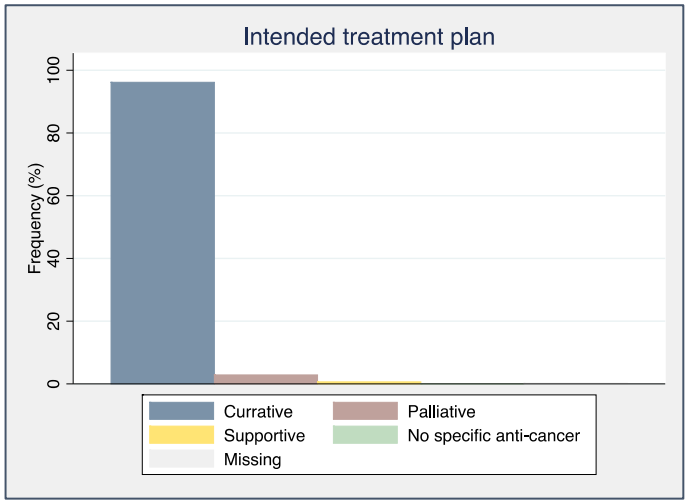
Figure 31: Bar chart of showing the comorbidity burden of participants in H&N5000.



5.3.2.6. Cancer care plan intent

Cancer care plan intent, as described in the baseline data collection section, refers to the intended course of treatment assigned at the time of diagnosis. The information needed to code this variable was obtained from the baseline DCF (section *B1*). The majority (96%; n=5,197/5,391) of participants were expected to be treated curatively (Figure 32).

Figure 32: Bar chart showing the intended treatment plan for participants at baseline in H&N5000.



5.3.3. Socioeconomic variables

5.3.3.1. Annual household income

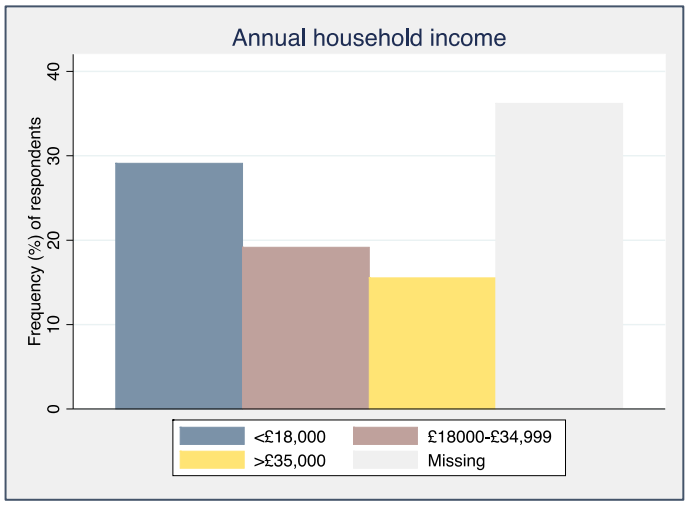
Participants' annual income was derived from question *A21* of the *About You* Questionnaire. Respondents were asked to describe their total household income from all sources, before tax and other deductions; they could select one of eight possible responses:

1) Less than £3,999;
2) £4,000 to £7,999;
3) £8,000 to £11,999;
4) £12,000 to £17,000;
5) £18,000 to £22,999;
6) £23,000 to £28,999;
7) 29,000 to £34,999;

8) £35,000 or more.

Responses were further categorised into three groups: <£18,000, £18,000 - £34,999 and >£35,000. Approaching half (46%; n=1,773/3,447) of respondents reported earning £18,000 or less (Figure 33).

Figure 33: Bar chart of participant income in H&N5000.



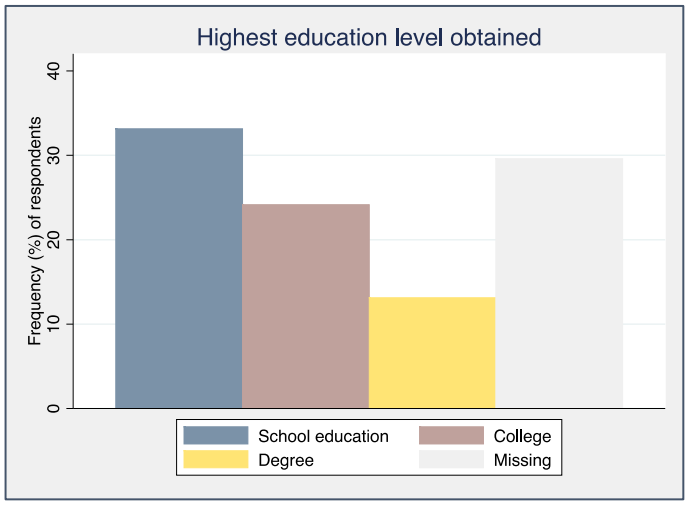
5.3.3.2. Highest education level obtained

In question *A7a* of the baseline *About You* questionnaire, participants' were asked: "What is the highest educational level you obtained?" The possible answers were:

1) Primary school;
2) Secondary school;
3) School or college sixth form;
4) College of Further Education;
5) Polytechnic or University;
6) Some other type of college

Alternatively, participants were given the option to specify their own attainment level themselves in a free-text box. Responses were collapsed into three categories, corresponding to "School education" (responses 1 and 2), "College" (responses 3, 4, 6), or "Degree" (response 5). Around half (47%; n=1,791/3,805) of respondents were educated to school-level (Figure 34).

Figure 34: Bar chart showing the education structure of respondents in H&N5000.



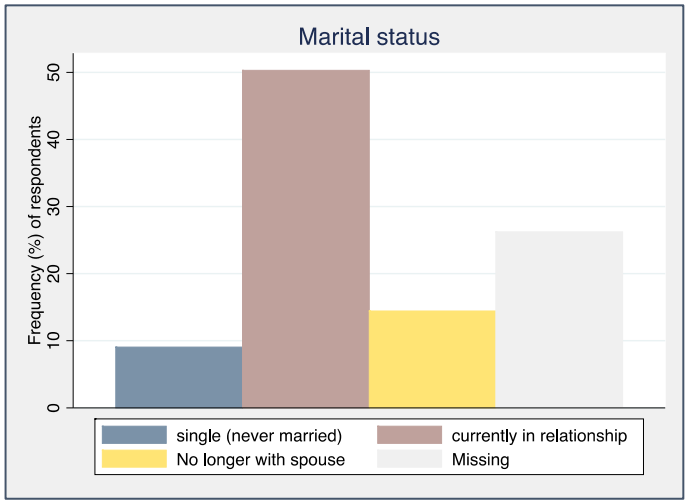
5.3.3.3. Marital status

Baseline marital status was determined from responses to question *A5* of the *About You* questionnaire. Participants were asked to indicate which of the six following statements best described their relationship status:

1) Single;
2) Widowed;
3) Separated;
4) Married;
5) Divorced
6) Living with a partner.

This information was coded as a categorical variable, as follows: “Single” (response 1), “Currently in a relationship” (responses 4 and 6) “No longer with spouse” (responses 2, 3, 5). The majority of participants (68%; n=2,718/3,987) were in a relationship at the time of enrolment into the study (Figure 35).

Figure 35: Bar chart showing H&N5000 participants' marital status.



5.3.4. Lifestyle behaviour variables

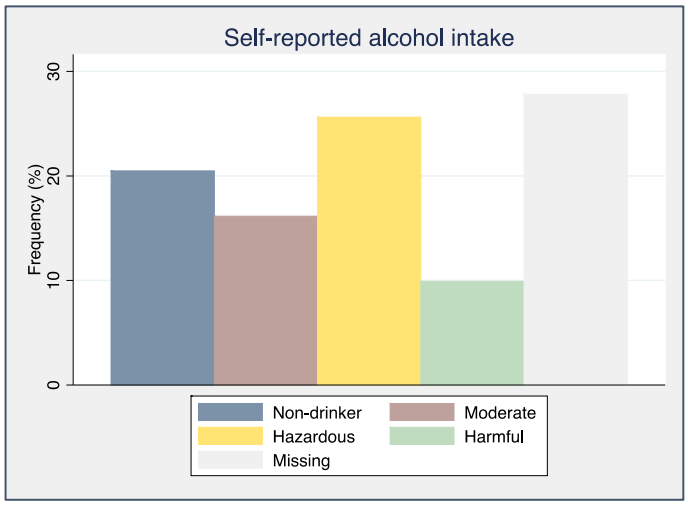
5.3.4.1. Alcohol consumption

Two core alcohol variables were derived based on information provided by participants in the baseline "*About You*" questionnaire. In question *A14*, respondents were asked: "Just before you became ill, how many alcoholic drinks did you have each week?". They were instructed to enter the number in the box. From these responses, a categorical variable was derived that provides the number of days per week (0-7) that alcohol was consumed before the individual became ill. In question *A15,* participants were asked: "About how many bottles of wine, spirits and pints of beer did you drink on average each week?" The question was divided into three parts corresponding to the three different alcoholic beverages-bottles of wine, bottles of spirits and bottles of beer, and respondents were asked to tick the box that best described their weekly intake. For example, in the bottles of wine column, participants could select one of seven boxes, ranging from 'None' to '11 or more'. Average alcohol intake was calculated based on the number of units of alcohol consumed, where one unit is measured as 10ml or 8g of pure alcohol [607].

In this thesis, four drinking categories were created based on UK guidelines (ref): none, moderate, hazardous and harmful. Moderate drinkers included men and women who drink < 14 units/week; hazardous drinkers included men who consumed 14 – 50 units/week and women who consumed 14 – 35 units/week; harmful drinkers included men who consumed > 50 units/week and women who consumed > 35 units/week. Participants were classed as non-drinkers if both drink-days per week was zero and units per week was zero. Overall,

28% (n=1,107/3,901), of respondents fell into the non-drinking category, 22% (n=873) were moderate drinkers, 36% (n= 1,385) were hazardous drinkers and 14% (n=536) were harmful drinkers (Figure 36).

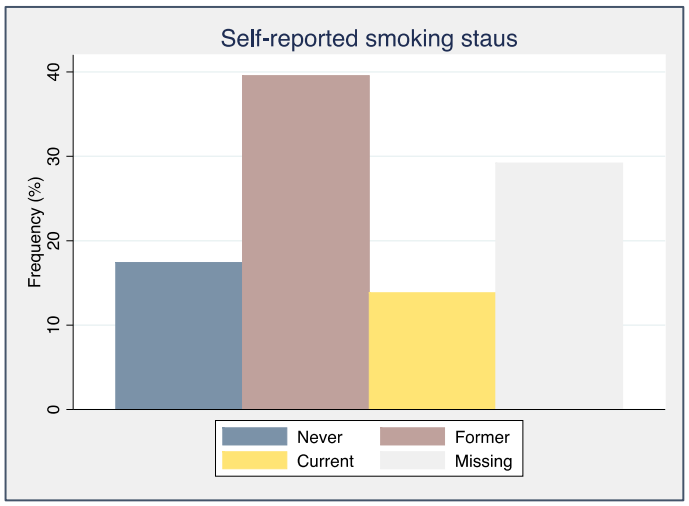
Figure 36: bar chart showing the alcohol consumption levels of participants in H&N5000.



5.3.4.2. Baseline smoking status

Baseline smoking status was based on participants' self-report. In Question *A8* of the "*About You*" questionnaire, participants were asked whether they would describe themselves as a "current", "former" or "never" user of tobacco. The definition of a "never-user", as outlined in the questionnaire, was someone who had never used tobacco on a regular basis, i.e. one tobacco product per day for a period of one year. By that definition, current smokers were defined as smokers of at least 1 cigarette per day over the course of a year at the time of blood sample collection. In total, 25% of participants (n=941/3,824) reported being never-smokers; 56% (n=2,138) were former smokers and 20% (n=747~) were current smokers (Figure 37).

Figure 37: Bar chart showing the proportion of never-, former- and current smokers in H&N5000.



5.4. Representativeness of the cohort

The ideal setting in which to carry out unbiased evaluations of associations between exposures and outcomes is a representative sample - one that is representative of the population in which you want to generalise your results to. The sample should be representative both in terms of confounders (to exposures and outcomes) and any other unmeasured variables that are not specified in the original study hypotheses.

H&N5000 is a UK-wide clinical cohort study. In population-based cohort studies, a sample of a defined population is selected for longitudinal assessment of exposure-outcome relations [608]. The extent to which a cohort sample is representative of the total population will depend on the completeness of the population frame available to the investigator (i.e. the proportion of the total population that is captured) and participation rates. If participation rates are low, this can challenge the interpretation of study results, both in terms of analytic and descriptive epidemiology.

As described early, H&N5000 is estimated to have captured a third of all incident cases in the UK, with approximately 50% of people identified as being eligible having been recruited. Whilst this represents the largest cohort of its kind in the UK currently, it is important to consider whether those individuals included in the study are typical of the HNC population as a whole. Given that the response rates of the 76 H&N5000 study centres varied from 20%-

90%, it is conceivable that certain geographic or sociodemographic areas may be over- or under-represented. The reasons why certain centres recruited fewer participants may be related to the centres themselves e.g. low motivation to partake, limited resources, or logistics, or to the individuals receiving care at those centres. If participation rates are low, this can challenge the interpretation of study results, both in terms of analytic and descriptive epidemiology.

From the histograms presented in this chapter, it is clear that, for the majority of variables, there was some missing data at baseline. The amount of missing data was less for variables that were recorded by research nurses on the data capture form e.g. tumour stage and comorbidity. For certain variables such as annual household income and smoking status, which were measured via the self-completed baseline questionnaire, the proportion of missing data is greater. If the people with missing data are not comparable to those without missing data, this could again influence the interpretation of study results. Missingness should therefore be considered in the study design.

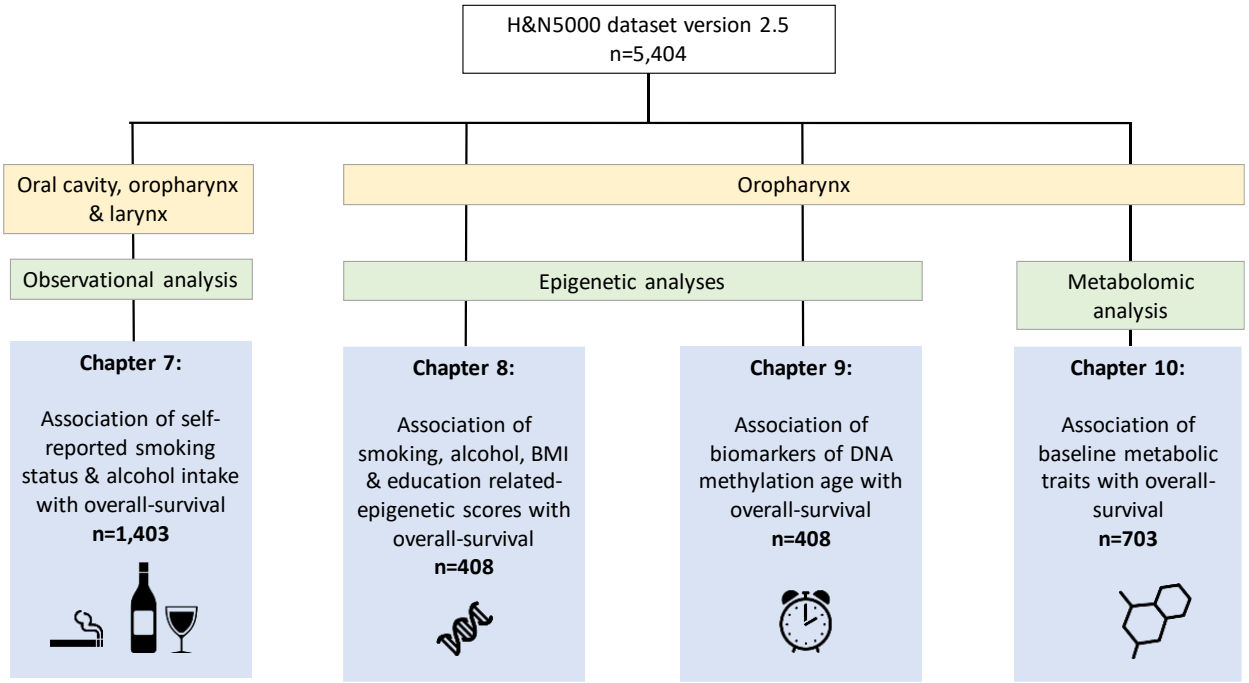
One positive feature of the prospective cohort study design is that individuals do not have the outcome of interest at the time of enrolment. Rather, the investigator will define the population of individuals to be included, measure the potential exposure of interest, and then follow the participants over time to evaluate for the occurrence of the outcome. This is an advantage because it limits selection bias into the study. Non-participation at follow- ups could introduce selection bias but because this thesis is interested in mortality risk and all H&N5000 participants were flagged with NHS digital at the start of the study for regular mortality updates, this is unlikely to be an issue here.

Chapter 6: A description of the datasets used in the thesis

6.1. Introduction

The H&N5000 dataset (version 2.5) comprises data on 5,404 participants. However, not all these individuals were included in the following analyses. The purpose of this chapter is to describe the baseline characteristics of the different analytic samples (primary analyses) used in this thesis. Values may differ in different releases of the H&N5000 dataset. The descriptive tables provide information on a range of clinical and sociodemographic variables but, for the purposes of analysis, some of these variables (e.g. IMD) may not have been included in the survival models. Full details of each of the models used will be provided in the relevant chapters.

Figure 38: number of participants included in each of the primary analysis conducted in this thesis.



"Cigarette" icon by Yeong Rong Kim, "Alcohol" icon by Aleksandr Vector, "DNA" icon by Milinda Courey, "Alarm clock" icon by kiddo, "Molecule" icon by Creative Stall from the Noun Project.

There are four separate results chapters in this thesis, as illustrated in Figure 38. The first chapter (Chapter 7) uses data on 1,403 individuals with cancers of the oral cavity,

oropharynx and larynx; the second and third chapters (chapters 8 and 9) use a dataset that includes 408 individuals with OPC, for whom epigenetic and genetic data are available; and the final results chapter (Chapter 10) uses a dataset of 703 individuals with OPC and metabolomic data available.

6.2. Participants included in the observational analysis

Table 19 describes the baseline characteristics of the 1,403 individuals included in the primary analysis, stratified by cancer site. In total, the sample population includes 404 individuals with oral cavity cancers, 656 with OPC and 343 with laryngeal cancer. There are differences in gender, age, tumour stage, HPV status, comorbidity, BMI, educational attainment, annual household income, deprivation level, relationship status and smoking status between tumour groups. There is no evidence of a difference in alcohol intake. The approximate ratio of males to females in each group is 3:2, 4:1 and 6:1 for oral cavity, oropharyngeal and laryngeal cancer, respectively. People with laryngeal cancer are generally older, with a mean age of 65 years (standard deviation [SD]=10.1) compared to 61 (SD=11.9) years for people with oral cavity cancers and 58 (SD=9.0) years for people with OPC. Mean BMI is highest among people with OPC (26.9 [SD=5.1] compared to 26.0 [SD=5.0] and 25.5 [SD=5.2] for people with oral cavity and laryngeal cancers). The proportion of current or former smokers in each group is 74%, 70% and 92% for oral cavity, oropharyngeal and laryngeal cancer, respectively. As expected, the proportion of HPV-positive individuals is substantially higher in OPC group (73% compared to 3% and 2% for oral cavity and laryngeal cancers, respectively).

The demographic and clinical attributes of people with HPV16 E6-seropositive and HPV-HPV16 E6 seronegative OPCs are compared in Table 20. People who were seropositive, hereafter referred to as HPV-positive, were more likely to have been diagnosed at an advanced tumour stage (75% were diagnosed at stage IV compared to 62% in the seronegative group), were less likely to be current smokers (6% versus 40%) and were less likely to drink hazardous to harmful amounts of alcohol (48% versus 62%). Groups were similar with respect to age at diagnosis (mean ages 59 and 58 for HPV-negative and HPV-positive individuals, respectively), gender and educational attainment but HPV-positive individuals had a higher BMI (28 versus 25) and were more likely to have an annual household income of over £35,000 (39% of people versus 23% in the HPV-negative group).

Table 19: Baseline characteristics of the study sample, stratified by tumour site (n=1,403).

	Oral cavity (n=404)		Oropharynx (n=656)		Larynx (n=343)		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*N*	*Frequency*	*p-value**
Gender							
Male	249	61.6%	533	81.3%	292	85.1%	
Female	155	38.4%	123	18.8%	51	14.9%	<0.001
TNM stage							
I	161	39.9%	21	3.2%	150	43.7%	
II	91	22.5%	73	11.1%	85	24.8%	
III	31	7.7%	86	13.1%	53	15.5%	
IV	121	30.0%	476	72.6%	55	16.0%	<0.001
HPV status							
Negative	392	97.0%	176	26.8%	336	98.0%	
Positive	12	3.0%	480	73.2%	7	2.0%	<0.001
Comorbidity							
None	179	44.3%	357	54.4%	136	39.7%	
Mild	137	33.9%	208	31.7%	121	35.3%	
Moderate/severe	88	21.8%	91	13.9%	86	25.1%	<0.001
Education							
School education	183	45.3%	278	42.4%	186	54.2%	
College	143	35.4%	247	37.7%	115	33.5%	
Degree	78	19.3%	131	20.0%	42	12.2%	0.003
Annual household income							
<£18,000	205	50.7%	225	34.3%	197	57.4%	
£18000-£34,999	112	27.7%	205	31.3%	95	27.7%	
>£35,000	87	21.5%	226	34.5%	51	14.9%	<0.001
IMD							
Low Deprivation	136	33.7%	219	33.4%	154	44.9%	
Moderate Deprivation	84	20.8%	151	23.0%	72	21.0%	
High Deprivation	184	45.5%	286	43.6%	117	34.1%	0.002
Relationship status							
single (never married)	56	13.9%	66	10.1%	43	12.5%	
currently in relationship	255	63.1%	484	73.8%	224	65.3%	
No longer with spouse	93	23.0%	106	16.2%	76	22.2%	0.003
Smoking status							
Never	107	26.5%	194	29.6%	29	8.5%	
Former	204	50.5%	362	55.2%	236	68.8%	
Current	93	23.0%	100	15.2%	78	22.7%	<0.001
Alcohol intake							
non-drinker	106	26.2%	159	24.2%	90	26.2%	
moderate drinker	86	21.3%	159	24.2%	75	21.9%	
hazardous-harmful drinker	212	52.5%	338	51.5%	178	51.9%	0.786
	N	*Mean (SD)*	*N*	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	404	61.1 (11.9)	656	58.3 (9.0)	343	65.3 (10.1)	<0.001
Body mass index (kg/m2)	404	26.0 (5.01)	656	26.9 (5.1)	343	25.5 (5.2)	0.024

** p-value for difference. Abbreviations: **TNM**, tumour, node and metastasis; **BMI**, body mass index; **HPV**, human papilloma virus; **IMD**, index of multiple deprivation*

Table 20: Baseline characteristics of the individuals with oropharyngeal tumours, stratified by HPV status (n=656).

	HPV negative (n=176)		HPV positive (n=480)		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*p-value**
Gender					
Male	137	77.8%	396	82.5%	
Female	39	22.2%	84	17.5%	0.176
TNM stage					
I	15	8.5%	6	1.3%	
II	30	17.0%	43	9.0%	
III	22	12.5%	64	13.3%	
IV	109	61.9%	367	76.5%	<0.001
Comorbidity					
None	78	44.3%	279	58.1%	
Mild	61	34.7%	147	30.6%	
Moderate/severe	37	21.0%	54	11.3%	0.001
Education					
School education	76	43.2%	202	42.1%	
College	67	38.1%	180	37.5%	
Degree	33	18.8%	98	20.4%	0.893
Annual household income					
<£18,000	86	48.9%	139	29.0%	
£18000-£34,999	50	28.4%	155	32.3%	
>£35,000	40	22.7%	186	38.8%	<0.001
IMD					
Low Deprivation	55	31.3%	164	34.2%	
Moderate Deprivation	46	26.1%	105	21.9%	
High Deprivation	75	42.6%	211	44.0%	0.497
Relationship status					
single (never married)	25	14.2%	41	8.5%	
currently in relationship	109	61.9%	375	78.1%	
No longer with spouse	42	23.9%	64	13.3%	<0.001
Smoking status					
Never	23	13.1%	171	35.6%	
Former	83	47.2%	279	58.1%	
Current	70	39.8%	30	6.3%	<0.001
Alcohol intake					
non-drinker	35	19.9%	124	25.8%	
moderate drinker	32	18.2%	127	26.5%	
hazardous-harmful drinker	109	61.9%	229	47.7%	0.005
	N	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	176	59.1 (9.5)	480	58.0 (8.8)	0.148
Body mass index (kg/m2)	176	24.9 (4.8)	480	27.6 (5.0)	<0.001

** p-value for difference. Abbreviations: **TNM**, tumour, node and metastasis; **BMI**, body mass index; **HPV**, human papilloma virus; **IMD**, index of multiple deprivation*

6.3. Participants included in the epigenetic analyses

The majority (78%) of individuals included in the epigenetic dataset are male (Table 21). Around 70% are HPV positive and over 85% were diagnosed with high stage tumours (TNM stages III or IV). The majority (71%) of individuals have a history of smoking (52% former smokers; 20% current smokers) and half reported drinking hazardous to harmful amounts of alcohol prior to getting ill.

Table 22 compares the characteristics of HPV-positive and HPV-negative individuals. There is evidence of a difference between groups with respect to age, tumour stage, comorbidity, BMI, annual household income, marital status, smoking status and alcohol intake but there is no apparent difference in gender, educational attainment or deprivation level. People with HPV-negative tumours are, on average, 3.3 years older than people with HPV-positive tumours. Compared to the HPV-positive group, the HPV-negative group are more likely to be former or current smokers (86% versus 66%) and are more likely to have drunk hazardous to harmful amounts of alcohol prior to their HNC diagnosis (61% versus 48%). The HPV-negative group are also more likely to have severe comorbidity (27% versus 15%) and less likely to be in the highest earning income group (23% earned ≥ £35,000 compared to 37% in the HPV-positive group).

Table 21: Baseline characteristics of all participants included in the epigenetic analyses (n=408).

Characteristic	*N*	*Frequency*
Gender		
Male	317	77.70%
Female	91	22.30%
TNM stage		
I	17	4.20%
II	39	9.60%
III	58	14.20%
IV	294	72.10%
HPV status		
Negative	122	29.90%
Positive	286	70.10%
Comorbidity		
None	211	52.10%
Mild	119	29.40%
Moderate/severe	75	18.50%
Education		
School education	170	43.70%
College	158	40.60%
Degree	61	15.70%
Annual household income		
<£18,000	138	38.70%
£18000-£34,999	103	28.90%
>£35,000	116	32.50%
IMD		
Low Deprivation	149	38.80%
Moderate Deprivation	81	21.10%
High Deprivation	154	40.10%
Relationship status		
single (never married)	47	11.70%
currently in relationship	280	69.70%
No longer with spouse	75	18.70%
Smoking status		
Never	110	28.10%
Former	205	52.30%
Current	77	19.60%
Alcohol intake		
non-drinker	104	26.00%
moderate drinker	90	22.50%
hazardous-harmful drinker	206	51.50%
	N	*Mean (SD)*
Age (years)	403	58.4 (9.6)
Body mass index (kg/m2)	272	26.4 (4.9)

** p- value for difference. Abbreviations: **TNM**, tumour, node and metastasis; **BMI**, body mass index; **HPV**, human papilloma virus; **IMD**, index of multiple deprivation.*

Table 22: baseline characteristics of participants included in the epigenetic analysis, stratified by HPV status (n=408).

	HPV- negative (n=122)		HPV- positive (n=286)		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*p-value*
Gender					
Male	96	78.70%	221	77.30%	
Female	26	21.30%	65	22.70%	0.753
TNM stage					
I	10	8.20%	7	2.40%	
II	18	14.80%	21	7.30%	
III	22	18.00%	36	12.60%	
IV	72	59.00%	222	77.60%	0.001
Comorbidity					
None	47	38.80%	164	57.70%	
Mild	41	33.90%	78	27.50%	
Moderate/severe	33	27.30%	42	14.80%	0.001
Education					
School education	52	45.60%	118	42.90%	
College	44	38.60%	114	41.50%	
Degree	18	15.80%	43	15.60%	0.861
Annual household income					
<£18,000	60	57.10%	78	31.00%	
£18000-£34,999	21	20.00%	82	32.50%	
>£35,000	24	22.90%	92	36.50%	<0.001
IMD					
Low Deprivation	51	42.90%	98	37.00%	
Moderate Deprivation	21	17.60%	60	22.60%	
High Deprivation	47	39.50%	107	40.40%	0.424
Relationship status					
single (never married)	21	17.80%	26	9.20%	
currently in relationship	62	52.50%	218	76.80%	
No longer with spouse	35	29.70%	40	14.10%	<0.001
Smoking status					
Never	17	14.40%	93	33.90%	
Former	50	42.40%	155	56.60%	
Current	51	43.20%	26	9.50%	<0.001
Alcohol intake					
non-drinker	30	25.20%	74	26.30%	
moderate drinker	17	14.30%	73	26.00%	
hazardous-harmful drinker	72	60.50%	134	47.70%	0.020
	N	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	122	60.7 (10.8)	79	57.4 (10.8)	0.002
Body mass index (kg/m2)	281	25.0 (4.8)	193	27.1 (4.9)	0.001

** p- value for difference. Abbreviations: **TNM**, tumour, node and metastasis; **BMI**, body mass index; **HPV**, human papilloma virus; **IMD**, index of multiple deprivation.*

6.4. Participants included in the metabolomics analysis

The primary metabolomics analysis, described in Chapter 10, included 703 individuals with OPC. The clinical and demographic characteristics of this analytic dataset are presented in Table 23.

There are around four times as many males than females included in this sample. The proportion of people with HPV-positive cancers is comparable to that of the epigenetic dataset at just over 70%. Similarly, the majority of these cancers (72%) were diagnosed at stage IV. Over half of the people with data reported being former smokers and 15% current smokers, whilst half drank hazardous to harmful amounts of alcohol.

Comparing the characteristics of people with and without HPV-driven tumours in this dataset (Table 24), there is evidence of a difference in age, stage, comorbidity, annual household income, relationship status and smoking and alcohol intake. HPV-positive individuals in this sample were younger at the time of diagnosis (around 58 years old compared to 60 years old), were diagnosed with higher stage tumours (77% were diagnosed with stage IV tumours compared to 62% in the HPV-negative group), were more likely to be in the highest income group (37% earned ≥ £35,000 compared to 22% of those in the HPV-negative group) and were more likely to be in a relationship (79% versus 60%). Only 6% of HPV-positive individuals in this sample population reported being current smokers at baseline compared to 40% in the HPV-negative group. The proportion of hazardous to harmful drinkers was 47% and 60% for HPV-negative and HPV-positive individuals, respectively.

Table 23: baseline characteristics of the study sample included in the metabolomic analysis (n=703).

Characteristic	*N*	*Frequency*
Gender		
Male	564	80.2%
Female	139	19.8%
TNM stage		
I	24	3.4%
II	73	10.4%
III	94	13.4%
IV	512	72.8%
HPV status		
Negative	185	26.3%
Positive	518	73.7%
Comorbidity		
None	379	53.9%
Mild	223	31.7%
Moderate/severe	101	14.4%
Education		
School education	293	41.7%
College	272	38.7%
Degree	138	19.6%
Annual household income		
<£18,000	249	35.4%
£18000-£34,999	222	31.6%
>£35,000	232	33.0%
IMD		
Low Deprivation	216	33.4%
Moderate Deprivation	149	23.0%
High Deprivation	282	43.6%
Relationship status		
single (never married)	68	9.7%
currently in relationship	519	73.8%
No longer with spouse	116	16.5%
Smoking status		
Never	216	30.7%
Former	382	54.3%
Current	105	14.9%
Alcohol intake		
non-drinker	179	25.5%
moderate drinker	171	24.3%
hazardous-harmful drinker	353	50.2%
	N	*Mean (SD)*
Age (years)	703	58.2 (9.1)
Body mass index (kg/m2)	703	27.0 (5.0)

** p-value for difference. Abbreviations: **TNM**, tumour, node and metastasis; **BMI**, body mass index; **HPV**, human papilloma virus; **IMD**, index of multiple deprivation.*

Table 24: baseline characteristics of the study sample, stratified by HPV status (n=703).

	HPV- negative (n=185)		**HPV- positive (n=518)**		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*p-value*
Gender					
Male	144	77.8%	420	81.1%	
Female	41	22.2%	98	18.9%	0.342
TNM stage					
I	18	9.7%	6	1.2%	
II	28	15.1%	45	8.7%	
III	25	13.5%	69	13.3%	
IV	114	61.6%	398	76.8%	<0.001
Comorbidity					
None	79	42.7%	300	57.9%	
Mild	64	34.6%	159	30.7%	
Moderate/severe	42	22.7%	59	11.4%	<0.001
Education					
School education	81	43.8%	212	40.9%	
College	71	38.4%	201	38.8%	
Degree	33	17.8%	105	20.3%	0.710
Annual household income					
<£18,000	95	51.4%	154	29.7%	
£18000-£34,999	50	27.0%	172	33.2%	
>£35,000	40	21.6%	192	37.1%	<0.001
IMD					
Low Deprivation	55	32.0%	161	33.9%	
Moderate Deprivation	43	25.0%	106	22.3%	
High Deprivation	74	43.0%	208	43.8%	0.759
Relationship status					
single (never married)	26	14.1%	42	8.1%	
currently in relationship	111	60.0%	408	78.8%	
No longer with spouse	48	25.9%	68	13.1%	<0.001
Smoking status					
Never	27	14.6%	189	36.5%	
Former	85	45.9%	297	57.3%	
Current	73	39.5%	32	6.2%	<0.001
Alcohol intake					
non-drinker	38	20.5%	141	27.2%	
moderate drinker	37	20.0%	134	25.9%	
hazardous-harmful drinker	110	59.5%	243	46.9%	0.014
	N	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	185	59.5 (9.7)	185	57.8 (8.8)	0.030
Body mass index (kg/m2)	518	25.1 (5.0)	518	27.7 (4.9)	<0.001

** p-value for difference. Abbreviations: **TNM**, tumour, node and metastasis; **BMI**, body mass index; **HPV**, human papilloma virus; **IMD**, index of multiple deprivation.*

6.5. Discussion

Version 2.5 of the H&N5000 database contains information on over 5,4000 individuals with HNC. This thesis focuses on a sub-sample of these participants. The number of individuals analysed at each stage varies because metabolomic and epigenetic data are only available for a subset of people with OPC, whilst clinical and demographic information has been collected on all consenting participants.

The purpose of this chapter was to introduce the three main datasets used in the primary analyses. The tables presented here demonstrate how, within datasets, there are differences in many of the clinical, socio-economic, and behavioural characteristics across tumour sites and between people with and without HPV-driven tumours. These differences may potentially confound any relationships that exists between the exposures and outcomes of interest, demonstrating the need to control for these factors in subsequent analyses.

The next chapter looks at the associations of self-reported smoking status and alcohol intake with survival in a group of participants with oral cavity, oropharyngeal and laryngeal cancer, after controlling for these clinical and lifestyle characteristics.

Chapter 7: Associations of self-reported smoking status and alcohol use at diagnosis with survival in H&N5000

This chapter includes sections from the publications below:

Beynon, R. A., Lang, S., Schimansky, S., Penfold, C. M., Waylen, A., Thomas, S. J., Pawlita, M., Tim, W., Martin, R. M., May, M. & , 2018. Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer**.** *International Journal of Cancer*. 143, 5, p.1114-1127**.**

7.1. Introduction

Earlier chapters have emphasized the importance of tobacco and alcohol use in HNC development. Although these behaviours together account for over 75% of new HNC cases [144], the prognostic significance of smoking status and alcohol intake at the time of HNC presentation remains unclear, especially for people with HPV-driven oropharyngeal tumours.

Most studies report a dose - dependent increase in mortality risk with increasing exposure to tobacco pre-diagnosis [441-443 445 450 455 456 459]. The magnitudes of the effects vary considerably however, as described in chapter 3. One possible explanation for the differences in the reported effect estimates is that studies have frequently been undertaken in single cancer sites, typically the larynx or oropharynx [450 455 456]. Where studies have included multiple sites, analyses have rarely stratified on this. In addition to this, studies have frequently been unable to adjust for important prognostic factors, such as comorbidity, BMI or HPV status, often because they were conducted retrospectively.

Evidence of an association between pre-treatment alcohol use and HNC mortality risk is conflicting. Some studies report an inverse association between alcohol intake and survival [408 442 450 609], whilst others have found little or no evidence of an effect [447 455]. Consequently, it is unclear whether any association of alcohol consumption with HNC cancer mortality is genuine, or the result of residual confounding by smoking or other factors. The results of one study suggest that the effects of alcohol intake on HNC survival may differ by treatment method and primary site [609], but this analysis only included 427 individuals from a single cancer center in Japan, emphasising the need for further research in this area.

The H&N5000 cohort provides a unique opportunity to examine the effects of smoking and alcohol drinking on HNC mortality owing to its large sample size, multi-center design, and the availability of detailed information on clinical, biological and lifestyle factors, including HPV status.

7.2. Aims and objectives

The overall aim of this chapter is to examine the effect of smoking status and alcohol intake, reported at the time of diagnosis, on survival in people with HNC in H&N5000. The specific objectives are:

1. to establish whether self-reported smoking status and alcohol consumption are associated with all-cause mortality in people with oral cavity, oropharyngeal and laryngeal cancer, after adjusting for important clinical, biological and socioeconomic factors;
2. to investigate whether any associations of smoking status and alcohol intake with all-cause mortality differ by tumour site;
3. to investigate whether any effects of smoking status and alcohol intake on all-cause mortality are influenced by HPV status (in OPC);
4. to explore the potential interactions of smoking and alcohol drinking and HPV status and smoking and/or alcohol drinking with all-cause mortality.

The results of an earlier analysis, which used version 2.3 of the H&N5000 data release, have been published in The International Journal of Cancer [575]. The analysis described here adopts the same statistical methods as the published analysis, but the sample numbers are slightly different, and the follow-up period is approximately one-year longer in the present analysis. In addition, this analysis is restricted to participants with SSC.

7.3. Methods

7.3.1. Study population

Included participants were those that:

- had cancers of the oral cavity, oropharynx and larynx (C01-C06; C09-C10; C32);
- had SCCs;
- had baseline questionnaire data and data capture available;

- were being treated with curative intent (at baseline);
- described their ethnicity as white.

Full details of the study including baseline data collection and follow-up have been provided in Chapter 5. To recap, participants were asked to complete three self-administered questionnaires at baseline (i.e. before their cancer treatment had commenced). The questionnaires enquired about social and economic circumstances, lifestyle behaviours, general health, and past sexual behaviours. Up-to-date treatment and cancer recurrence information was extracted from participants' medical records and entered into a data capture form.

7.3.2. Outcome assessment

The outcome of interest was all-cause mortality. Follow-up for survival analysis was defined as the time in years from study enrolment to the first of: date of death from any cause or the date of censorship (mortality records linked up until 11/10/2018). Participants were censored if they have not been observed long enough for the event to occur, i.e. the dataset closed before the event could be observed, or they were lost to follow-up. Given that participants were flagged with NHS digital (for mortality updates), the number of participants loss to follow-up is expected to be low in this analysis.

7.3.3. Defining exposures

Information on tobacco and alcohol history was obtained at baseline via the self-administered questionnaire. Full details are described in Chapter 5. Briefly, participants were asked to report their current smoking status ("never", "former" or "current") and the number of bottles of wine, spirits, and pints of beer they drank on an average week (prior to their HNC). Baseline drinking categories were then defined as "none", "moderate", or "hazardous to harmful". Smoking categories were preserved to allow comparison with the published literature, wherein most authors have adopted these classifications [441-443 445 450 451 456 457 459]. Alcohol consumption categories were derived based on current UK-drinking guidelines[607] , to provide clinical ease of interpretation.

7.3.4. Assessment of HPV status

Plasma was analysed for antibodies against the HPV16 E6 oncoprotein, using a median MFI cut-off of ≥1,000 MFI (See chapter 5 for details). Participants with values below this cut-off were classed as HPV16 E6- seronegative, hereafter referred to as "HPV-negative", and

those with values above the cut off were defined as HPV16 E6 seropositive, hereafter “HPV-positive”.

7.3.5. Statistical analysis

Analyses were completed using Stata version 15.1. (StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP).

7.3.5.1. Descriptive analysis

Baseline descriptive data were stratified by tumour site and HPV status (oropharyngeal only). *P*- values were calculated with the $\chi 2$ test for categorical variables and analysis of variance (ANOVA) for continuous variables.

7.3.5.2. Accounting for missing data

Missing data is a common problem in epidemiologic studies, particularly for research using questionnaire data. There are several ways of dealing with missing data. Commonly, complete-case analyses are performed, whereby data is only analysed on the individuals who had data recorded on all variables of interest. Depending on the mechanism of missing data however, this approach can produce biased estimates [610]. That is, if the remaining subjects are not representative of the original population, then the results will be misleading. Even if the results are not biased, the analysis is inefficient because potentially useable data is ignored. As such, complete case analyses may lead to a loss of precision and power [610-612].

An alternative approach is to impute missing data. Multiple imputation (MI) involves generating a set of plausible replacements for each missing value (using the correlations in the observed dataset), running the analysis in each imputed dataset, and then combining the point estimates in a way that reflects the uncertainty around the true value [610]. The correct number of new ‘complete’ datasets to be generated, i.e. the number of imputations to be performed (*m*), is an area of debate, but it is often stated that *m* should be at least 5 [610]. An appropriate *m* will however depend on the amount of missing data being imputed. Overall, the validity of the resulting analysis depends on the assumptions made about the ‘missing mechanism’. If data are missing not at random (MNAR), that is if missingness is related to the value of the missing observation, then MI will not solve the problem. For instance, BMI would be MNAR if a participant was less likely to report their weight if they knew they were overweight. By contrast, if data were missing completely at random (MCAR) or missing at random (MAR), i.e. data are missing purely by chance or missingness is related to other

factors that are recorded within the dataset and not on unobserved values, then MI may increase the efficiency and precision of the analysis.

In the current analysis, BMI was missing for a large proportion of participants (further details on the proportion of missingness below) because information on height and weight were not collected at the start of the study. It was therefore assumed that the missingness mechanism for BMI was MAR, since missingness was dependent on the date of enrolment. Several other variables had missing data. The missing status of each of these variables was considered using the Stata command `mvpatterns`, which produces a table showing how frequently missing values in each of the variables occur together. Missing values were imputed using the `ice` package for multiple chained equations in Stata [613]. Twenty imputed datasets were generated for each of the three tumour sites and combined using Rubin's rule to obtain valid statistical inferences [614]. The three resulting datasets were merged into one single analytical dataset. Imputation models included all of the variables in the substantive Cox model, the event status (i.e. a binary variable indicating whether the participant experienced the event (death) or not), and the Nelson-Aalen estimator of the cumulative hazard [615].

7.3.5.3. Survival analysis

Kaplan-Meier survival curves were first plotted to visualize differences in survival between tumour sites in relation to smoking and alcohol status. The log-rank test was used to test the null hypothesis that there was no difference between groups with respect to the probability of the event (death) at any time point. The variance in survival explained by smoking and alcohol intake was calculated via the `str2d` Stata module using the method of Royston and Sauerbrei [616]. The explained variation statistic (R_2) is based on their index of discrimination (D) for proportional hazards for censored survival data.

The primary analyses included complete cases only i.e. participants with complete data for confounders used in the adjusted models and information on smoking and alcohol consumption (as per the published manuscript). Cox proportional hazards models, stratified by tumour site, stage and HPV status, were used to examine the associations of baseline smoking status and alcohol intake with survival. Only oropharyngeal cases were considered in the HPV stratified models because the role of HPV in tumours outside the oropharynx is uncertain, as is the ability of serology to detect HPV driven tumours in other anatomical sites. HRs and 95% CIs for mortality were calculated for each category of smoking and drinking, using never-smokers and non-drinkers as the reference groups.

The equation for the Cox model can be written as:

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)$$

Where h(t) is the hazard function, i.e., the probability of having the event (death) at time *t* given the individual survived to time *t*, $h_0(t)$ is the baseline hazard, i.e., the hazard for the respective individual when all covariate values are equal to zero, and $\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$ represents the covariate effects. The predicted hazard (i.e., h(*t*)), is therefore the product of the baseline hazard ($h_0(t)$) and the exponential function of the linear combination of the predictors.

A major assumption of the Cox proportional hazards (PH) model is that hazards between levels of covariates are constant, or proportional, over time. This means that if a person's risk of dying at some initial time point is twice as high as that for another person, their risk of death remains twice as high at all other time points. This assumption of proportional hazards should be tested. In the current analysis, The PH assumption was checked by plotting the log of the cumulative hazard against time (log-log plots) and checking for parallelism and using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals [617] (implemented using the "`estat phtest`" command in Stata). The PH assumption is violated when there is a significant relationship between residuals and time.

Four different Cox models were fitted:

1) a minimally adjusted model that included age and gender;
2) a model that additionally adjusted for clinical factors (TNM stage, HPV status, BMI and comorbidity);
3) a model that additionally adjusted for socioeconomic factors (annual household income, education level and marital status);
4) a fully adjusted model that included both smoking and alcohol intake, in addition to clinical and socioeconomic factors.

The covariates were selected on the basis of the strength of prior evidence linking them with HNC survival (see chapter 2).

Potential interactions between tumour stage and smoking, tumour stage and alcohol consumption, HPV status and smoking, HPV status and alcohol consumption and smoking and alcohol intake were investigated by fitting an interaction term in the models and using a likelihood ratio test. As above, HPV analyses were restricted to the subset of participants with oropharyngeal tumours.

To control for the fact that people treated in some HNC centres may be more likely to die than others for unobserved reasons not captured by the model covariates, a 'shared frailty' term was fitted (with gamma distribution) to model 4 (the fully adjusted model) as a sensitivity analysis, to assess for the presence of heterogeneity between recruitment centres. The significance of the frailty component was tested using a likelihood-ratio test. Finally, the primary analysis was repeated in the imputed dataset. All reported *p*-values are two-sided.

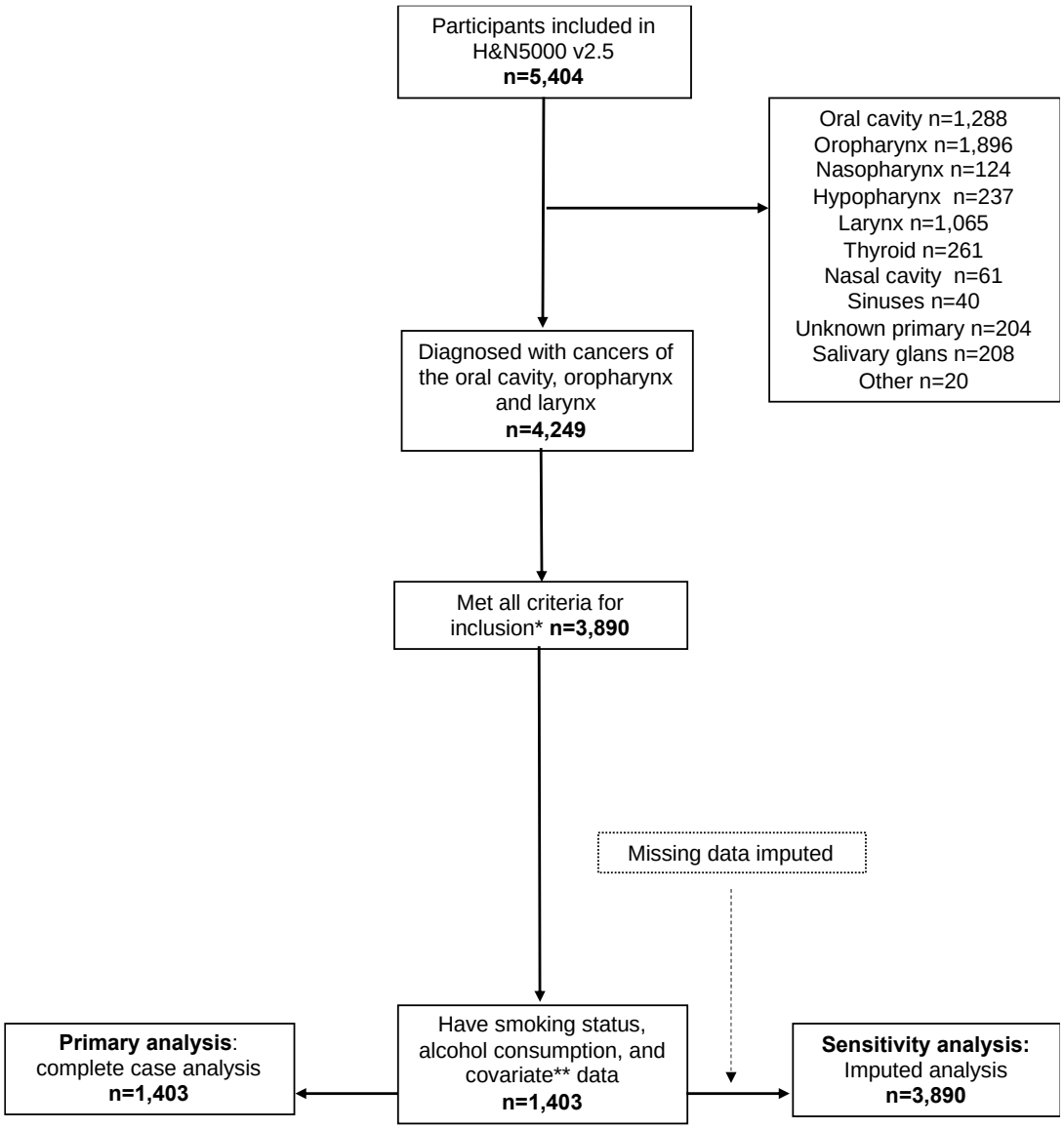
7.4. Results

The dataset included 5,404 people with HNC (Figure 39). Of those, 3,890 had cancers of the oral cavity, oropharynx and larynx and were eligible for inclusion in the analysis (i.e. they had SCC, were being treated with curative intent and were white). The complete case analysis included 1,403 individuals. Missing covariate data was imputed for 2,487 participants.

7.4.1. Missing data

The distribution of missing data stratified by tumour site and HPV status is shown in Table 25. Overall, the largest proportion of missing data was present for BMI (41%), followed by annual household income (36%) and education (29%). Smoking and drinking data were missing for 28% and 27% of participants, respectively. The proportion of missing data was comparable across tumour sites.

Figure 39: Flow of Head and Neck 5000 participants through the study.



** Participants with SCC, being treated with curative intent, and who describe their ethnicity as white. ** Age, gender, TNM stage, HPV status, BMI, comorbidity, education, annual household income, index of multiple deprivation (IMD), smoking status and alcohol intake.*

Table 25: Proportion of missing data, overall and stratified by tumour site.

	All sites (N=3,890)	Oral cavity (N=1,150)	Larynx (N=970)	Oropharynx		
				All (N=1,770)	HPV (+) (N=455)	HPV (-) (N=1,032)
Variable	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N (%)**	**N (%)**
Age	23 (0.6)	11 (1)	1 (0.1)	11 (0.6)	1 (0.2)	8 (0.8)
Gender	0	0	0	0	0	0
TNM stage	10 (0.3)	3 (0.3)	0	7 (0.4)	1 (0.2)	5 (0.5)
HPV status	608 (15.6)	174 (15.1)	151 (15.6)	283 (16.0)	0	0
Comorbidity	74 (1.9)	23 (2.0)	22 (2.3)	29 (1.6)	7 (1.5)	15 (1.5)
BMI	1599 (41.1)	480 (41.7)	381 (39.3)	738 (41.7)	208 (45.7)	389 (37.7)
Household income	1383 (35.6)	430 (37.4)	362 (37.3)	591 (33.4)	181 (39.8)	290 (28.1)
IMD	367 (9.4)	118 (10.2)	98 (10.1)	151 (8.5)	86 (8.3)	31 (6.8)
Education	1133 (29.1)	331 (28.8)	300 (30.9)	502 (28.4)	152 (33.4)	241 (23.4)
Marital status	996 (25.6)	283 (24.6)	254 (26.2)	459 (25.9)	140 (30.8)	215 (20.8)
Smoking status	1101 (28.3)	318 (27.7)	278 (28.7)	505 (28.5)	145 (31.9)	251 (24.3)
Alcohol intake	1061 (27.3)	309 (26.9)	274 (28.3)	478 (27.0)	144 (31.7)	227 (22.0)

*Abbreviations: **BMI**, body mass index; **HPV**, human papillomavirus; **IMD**, index of multiple deprivation; **n**, number; **%** , percent missing.*

7.4.2. Baseline characteristics of study population

The demographic and clinical attributes of the 1,403 individuals (oral cavity, n=404; oropharynx, n=656; larynx, n=343) included in the complete case analysis were presented in Chapter 4 (Table 19 by tumour site; Table 20 by HPV status). When considering similarities and differences across tumour sites, groups were different with respect to all of the characteristics considered, except for alcohol intake, where approximately 50% of people in each group drank hazardous to harmful amounts of alcohol. People with laryngeal cancers tended to be older overall and were more likely to be male and were more likely to be current smokers. People with HPV-positive OPC were more likely to have high stage tumours (stages III and IV) and were generally more affluent than people with HPV-negative cancers.

The baseline descriptives of individuals included in the imputed analysis (n=3,890) are presented in the supplementary material (Appendix 1). As was the case with the complete case dataset, there were differences across tumour groups in all variables except alcohol intake (p=0.660). The ratio of males to females in each tumour group was comparable to the complete case dataset, as were the mean ages of participants (oral cavity, 62 years [SD=12.3]; oropharynx, 59 years [SD=9.0]; larynx, 65 years [SD=10.4]) and the proportion of former and current smokers (oral cavity, 75%; oropharynx, 73%; larynx, 81%).

Differences between the sample of people with complete data and those with missing data are compared in Table 26. There were differences between the two groups with respect to the burden comorbidity (p-value for difference <0.001), HPV status (p=0.009), educational level (p=0.052), income (p=0.031) and deprivation level (p<0.001). The sample of people with missing data were more likely to: have higher comorbidity, be HPV-negative, be lower educated, and be in a lower income group. However, people with missing data were less likely to live in the lowest deprivation group than people with complete data. It is important to note that IMD index was not available for Scotland. There was no evidence to suggest that people with missing data differed from those with complete data with respect to age, sex, tumour stage, BMI, marital status, smoking status, or alcohol intake.

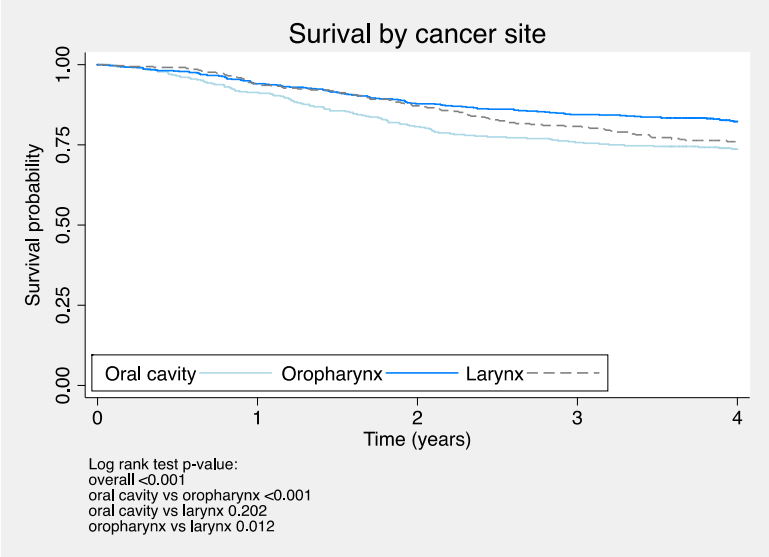
Table 26: A comparison of the baseline characteristics of participants who did and did not have missing data.

	Missing data (n=2,487)		Complete data (n=1,403)		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*p-value*
Tumour site					
Oral Cavity	746	30.00%	404	28.80%	
Oropharynx	1114	44.80%	656	46.80%	
Larynx	627	25.20%	343	24.40%	0.495
Gender					
Male	1868	75.10%	1074	76.60%	
Female	619	24.90%	329	23.40%	0.315
TNM stage					
I	532	21.50%	332	23.70%	
II	439	17.70%	249	17.70%	
III	337	13.60%	170	12.10%	
IV	1169	47.20%	652	46.50%	0.316
HPV serology group					
HPV-negative	1292	68.80%	904	64.40%	
HPV-positive	587	31.20%	499	35.60%	0.009
Comorbidity					
None	941	39.00%	672	47.90%	
Mild	877	36.30%	466	33.20%	
Moderate/Severe	595	24.70%	265	18.90%	<0.001
Education level					
School education	686	50.70%	647	46.10%	
College	439	32.40%	505	36.00%	
Degree	229	16.90%	251	17.90%	0.052
Annual household income					
<£18,000	549	49.70%	627	44.70%	
£18000-£34,999	307	27.80%	412	29.40%	
>£35,000	248	22.50%	364	25.90%	0.031
IMD					
Low Deprivation	936	44.20%	509	36.30%	
Moderate Deprivation	459	21.70%	307	21.90%	
High Deprivation	725	34.20%	587	41.80%	<0.001
Relationship status					
Single (never married)	188	12.60%	165	11.80%	
Currently in relationship	998	66.90%	963	68.60%	
No longer with spouse	305	20.50%	275	19.60%	0.607
Smoking status					
Never	291	21.00%	330	23.50%	
Former	815	58.80%	802	57.20%	
Current	280	20.20%	271	19.30%	0.273
Alcohol consumption					
Non-drinker	377	26.40%	355	25.30%	
Moderate drinker	292	20.50%	320	22.80%	
Hazardous-harmful drinker	757	53.10%	728	51.90%	0.313
	N	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	2464	61.76 (10.84)	1403	60.80 (10.54)	0.241
BMI	888	26.35 (5.32)	1403	26.55 (5.12)	0.204

7.4.3. Kaplan-Meier survival plots

There were 329 deaths (oral cavity, n=118 [29%]; oropharynx, n=119 [19%]; larynx, n=89 [26%]) during a median follow-up time of 4.3 years (IQR=3.7 years - 5.1 years). Overall, there was a difference in the probability of survival across tumour sites (p<0.001; Figure 40). On visual inspection of the Kaplan-Meier plot, the survival probability was lowest for people with oral cavity cancers. Up to 2-years post diagnosis, there was little difference in survival probabilities for OPC and laryngeal cancer cases but after 2-years, the cumulative survival probability for people with laryngeal cancer was lower than for people with OPC.

Figure 40: Kaplan-Meier plot of overall survival by HNC site.



As illustrated in Figure 41, the risk of death increased with increasing smoking exposure in all cancer groups (log-rank *p*-value across all three smoking categories: oral cavity, *p*<0.001; oropharynx, *p*=0.001; larynx, *p*=0.001). For oral cavity cancers, the risk of death (post 2-years) appeared to be similar for former and current smokers. Considering alcohol exposure, there was evidence of a difference in the probability of survival across categories of intake in the oral cavity and OPC groups (log-rank *p*-value across all three alcohol categories: oral cavity, *p*=0.013; oropharynx, *p*=0.066). Kaplan-Meier plots suggest that moderate drinkers had the lowest mortality risk. The amount of alcohol consumed did not appear to influence survival in the laryngeal cancer group (log-rank *p*-value across all three alcohol categories: *p*=0.577). When plots were stratified by HPV status (Figure 42), there was a trend for increasing mortality risk with increasing tobacco exposure in the HPV-negative OPC group (log-rank *p*-value across all three smoking categories*: p*=0.018). In the HPV-positive group,

current smokers appeared to have a higher mortality risk compared with never-smokers (log-rank *p*-value for never vs. current smokers: 0.076). There was evidence of a difference in survival probability across alcohol categories in the HPV-positive group (log-rank *p*-values: across all three alcohol categories $p=0.013$; for non-drinkers vs hazardous-harmful drinkers $p=0.044$) but not the HPV-negative group (log-rank *p*-value across all three alcohol categories: $p=0.788$).

Figure 41: Kaplan-Meier plot of overall survival by smoking status and alcohol consumption, stratified by tumour site.

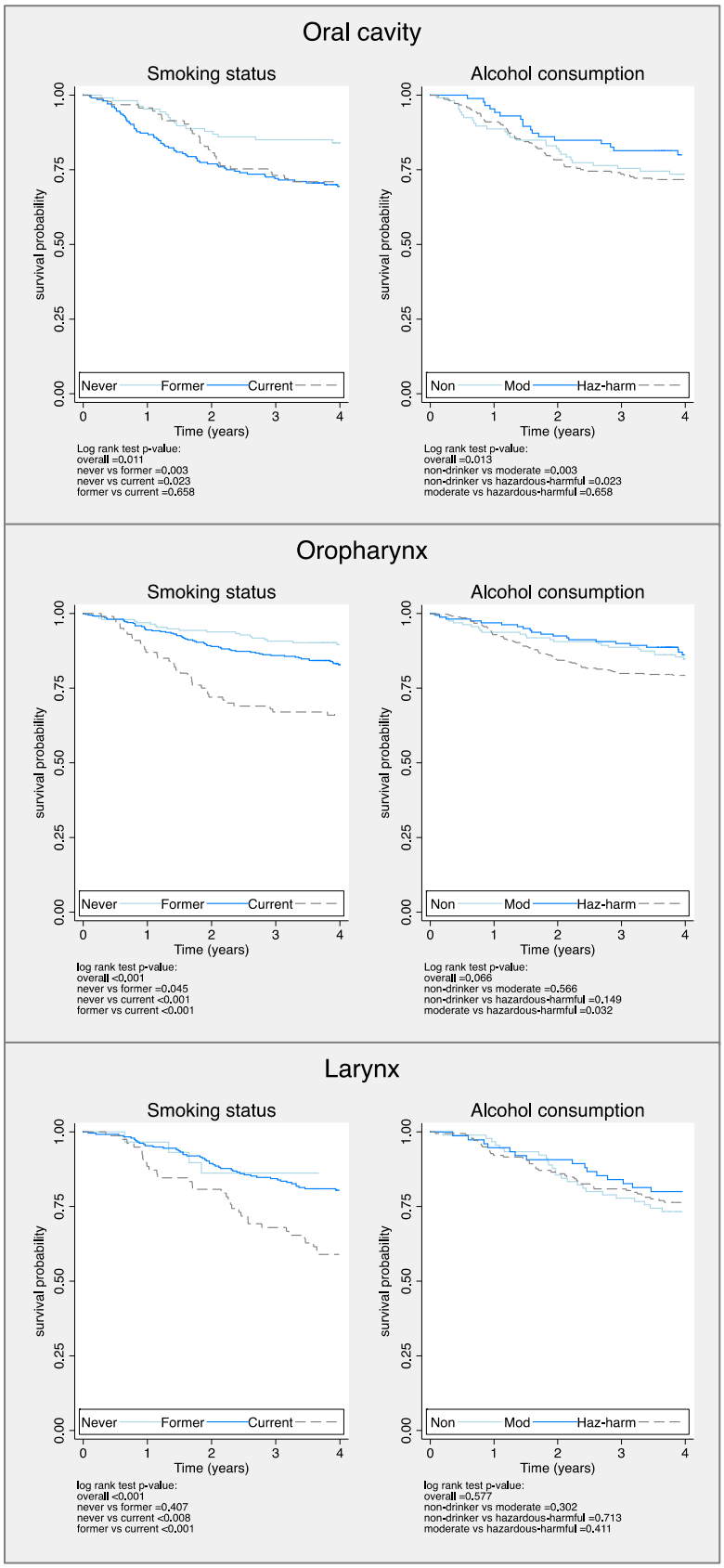
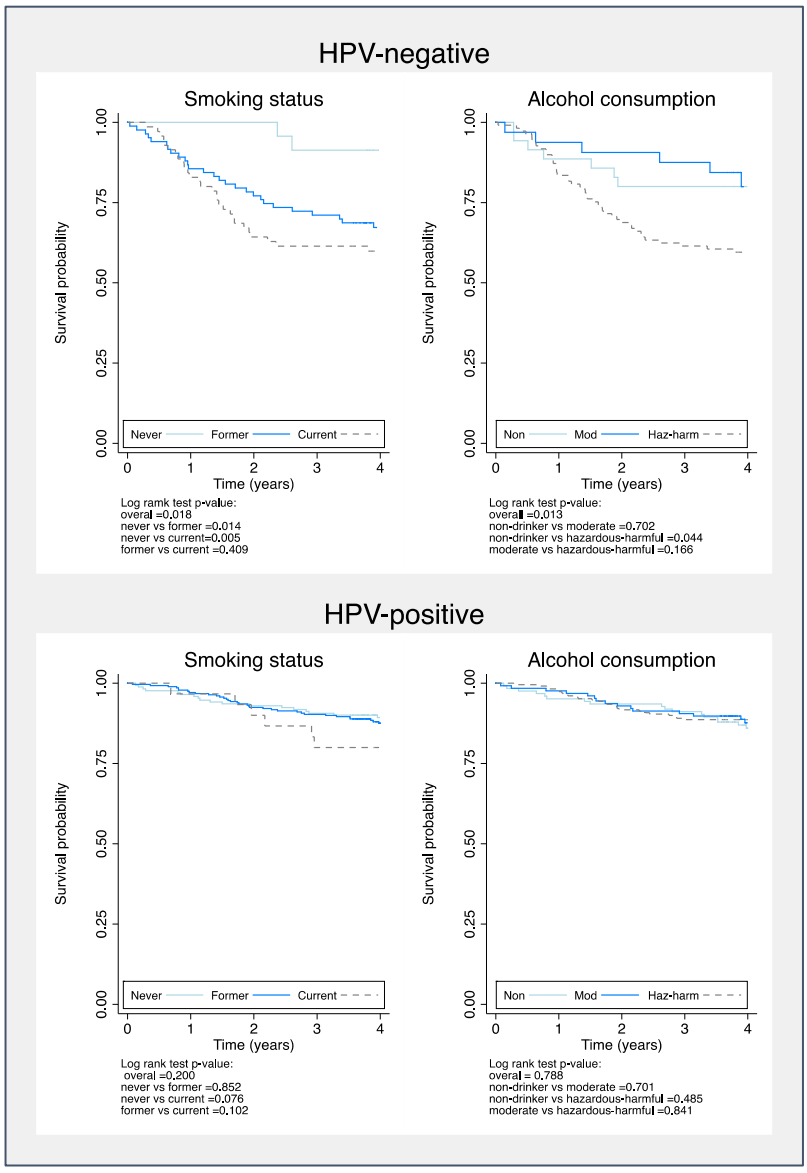


Figure 42: Kaplan-Meier plot of overall survival by smoking status and alcohol consumption, stratified by HPV status.



7.4.4. Variation in survival explained by smoking status and alcohol intake

A minimally adjusted model that included age and gender accounted for 6% of the variation in survival (R_2=0.06 [95% CI:0.02, 0.10]). Inclusion of self-reported smoking status led to an increase in the proportion of explained variation of 7% (R_2=0.13 [95% CI:0.08, 0.19]). Adding self-reported alcohol intake to the model led to an increase of 2% (R_2=0.08 [95% CI:0.04, 0.12]).

7.4.5. Smoking status and survival

For all cancer sites combined (n=1,403), there was strong evidence of an association between smoking status at diagnosis and survival. Compared to never smokers, the HR for current smokers was 3.3 (95% CI: 2.3, 4.7; *p* for trend <0.001) in the minimally adjusted model; this attenuated to 2.0 (95% CI: 1.4, 3.0; *p* <0.001) in the fully adjusted model (Table 27). For former smokers, the HR was 1.7 (95% CI: 1.2, 2.3; *p* <0.001) when models were minimally adjusted and 1.6 (95% CI: 1.2, 2.3; *p* = 0.001) on full adjustment.

In stratified analyses, the associations of smoking status with survival were present across all tumour groups in the minimally adjusted models, but only the oral cavity and laryngeal cancer groups were robust to adjustment (Table 27). In the oral cavity cancer group, the HR for current smokers was 2.2 (95% CI: 1.2, 3.8; *p* = 0.012) in the minimally adjusted model and 2.3 (95% CI: 1.2, 4.4; *p* = 0.010) in the fully adjusted model for people with laryngeal cancers, the respective HRs were 4.4 (95% CI: 1.5, 12.4; *p* <0.001) and 3.4 (95% CI: 1.1, 10.3; *p* for trend =0.001).

Results of the imputed analysis (n=3,890; number of deaths: oral cavity, n=378/1,150; oropharynx, n=425/1,770; larynx=268/970) were comparable to those of the complete case analysis (Appendix A2). In minimally adjusted models, the hazard ratio was 3.2 (95% CI: 2.6, 4.0) and 1.7 (95% CI: 1.3, 2.1; *p* <0.001) for current and former smokers, respectively. In fully adjusted models, the corresponding hazard ratios were 1.9 (95% CI: 1.5, 2.5) and 1.5 (95% CI: 1.2, 1.9; *p* for trend <0.001). In the stratified analysis, the associations of smoking status with survival were robust to adjustment in all tumour groups (HR for former vs. never smokers = 1.7 (95% CI: 1.1, 2.6; *p* = 0.006); 1.9 (95% CI: 1.2, 3.1; *p* = 0.004) and 3.1 (95% CI: 1.5, 6.4; *p* <0.001) for oral, oropharyngeal and laryngeal cancers, respectively).

7.4.6. Alcohol intake and survival

In the minimally adjusted model, the HR for hazardous to harmful drinkers compared to non-drinkers was 1.2 (95% CI: 1.0, 1.6; *p* = 0.042) (Table 27). On full adjustment, there was no evidence for an increase in mortality risk (HR= 1.0 [95% CI: 0.8, 1.3; *p* = 0.554]). There was a suggestion that moderate drinkers experienced improved survival compared to non-drinkers (minimally adjusted HR= 0.7 [95% CI: 0.5, 1.1; *p* = 0.042]; fully adjusted HR= 0.8 [95% [CI: 0.5, 1.1; *p* = 0.554]).

When models were stratified by tumour site, there was little evidence of an association between alcohol consumption and survival (Table 27). Hazard ratios for hazardous to harmful drinkers were 0.92 (95% CI: 0.58, 1.45; *p* = 0.789), 1.28 (95% CI: 0.80, 2.05; *p* = 0.263) and 0.94 (95% CI: 0.55, 1.59; *p* = 0.798) for oral cavity, oropharyngeal and laryngeal cancers, respectively.

In the imputed analyses (Appendix A2), a comparable pattern of association between alcohol drinking and survival was observed (minimally adjusted HRs: 1.2 [95% CI: 1.0, 1.4; *p* = 0.012] and 0.9 [95% CI: 0.7, 1.1] for hazardous/harmful drinkers and moderate drinkers respectively; corresponding fully adjusted HRs:1.1 [95% CI: 0.9, 1.3; *p* = 0.321] and 0.9 [95% CI: 0.7, 1.1]). As was the case in the primary analysis, there was little evidence that alcohol consumption influenced mortality risk in the stratified analysis.

7.4.7. Heterogeneity between centres

Unobserved heterogeneity by recruitment centre was examined using a shared frailty model, wherein the random effect is common or "shared" by people in the same centre or "cluster" [618]. The likelihood ratio test of theta (θ) =0 (testing the null hypothesis that the variance component in the variable "centre" is not different from zero) gave a chi-squared of 0.46 with an estimated *p*-value of 0.250, suggesting there is little variance in survival time among centres.

7.4.8. Influence of tumour stage on the associations of smoking and alcohol intake with survival

There were 581 people with low-stage tumours (stages I-II) and 822 with high-stage tumours (III-IV) in the analysis. Of those, 109 people in the low stage group and 220 in the high stage group died during the follow-up period. In minimally adjusted models (Table 28), the HR of death for current versus never-smokers was 4.0 (95% CI: 1.9, 8.3; *p*< 0.001) in the low-stage group and 3.6 (95% CI: 2.4, 5.4; *p* <0.001) in the high-stage group. On full adjustment, the corresponding HRs were 2.8 in the low stage group (95% CI: 1.3, 6.2; *p* = 0.011), and 1.9 in the high stage group (95% CI: 1.2, 3.0; *p* <0.003). The point estimates for moderate-drinkers in the low-stage group were similar after adjustment (minimally adjusted HR= 0.80 [95% CI: 0.90, 2.32]; fully adjusted HR= 0.78 [95% CI: 0.42, 1.44]). For hazardous/harmful drinkers, the HR attenuated slightly (minimally adjusted HR=1.44 [95% CI: 0.90, 2.32]; fully adjusted HR=1.30 [95% CI: 0.81, 2.09]). In the high-stage group, the HR for moderate-

drinkers was 0.66 in the minimally adjusted model (95% CI: 0.44, 0.99) and 0.73 in the fully adjusted model (95% CI: 0.48, 1.11). The corresponding HRs for hazardous/harmful drinkers were 1.10 (95% CI: 0.80, 1.52) and 0.97 (95% CI: 0.70, 1.35).

Comparable results were found in the imputed analysis (Appendix A3). In the low stage tumour group, current smokers had a 2.6-fold increased risk of death compared to never smokers (fully adjusted HR=2.6 (95% CI: 1.6, 4.2; *p*<0.001), whilst former smokers were 80% more likely to die during follow-up (fully adjusted HR=1.8 [95% CI: 1.1, 2.9]. Among the high stage tumour group, current and former smokers had a 70% and 30% increased risk of death, respectively (fully adjusted HR=1.7 [95% CI: 1.3,2.3; p<0.001] and 1.3 [95% CI: 1.0, 1.7]). As was the case in the primary analysis, alcohol intake was not associated with survival in either group, after adjusting for clinical factors.

7.4.9. Influence of HPV status on the associations of smoking and alcohol intake with survival

In total, 176 of the 656 OPC cases were HPV negative (Table 29). There were 61 deaths in each group. There was limited evidence of an association of smoking with survival when models were stratified (possibly due to limited statistical power). In the HPV-negative group, smoking status was associated with survival in the minimally adjusted model, but CIs were wide (HR=6.9 [95% CI: 1.6, 29.4] for current smokers and HR=5.21 [95% CI: 1.24, 21.96; *p*=0.005]) for former smokers. There was some attenuation on adjustment (fully adjusted HRs of 2.66 [95% CI: 0.55, 16. 12.80] and 3.70 [95% CI: 0.82, 16.78; p=0.492] for current and former smokers, respectively). The effect estimates and corresponding CIs were smaller in the HPV-positive group (minimally adjusted HRs of 2.18 [95% CI: 0.91, 5.19] and 1.00 [95% CI: 0.57, 1.74; p=0.272] for current and former smokers; fully adjusted HRs of 2.13 [95% CI: 0.84, 5.14] and 1.04 [95% CI: 0.59, 1.84; *p*=0.311] for current and former smokers).

Considering alcohol consumption and survival, there was weak evidence of an association in the HPV-negative group (minimally adjusted HRs of 0.82 [95% CI: 0.28, 2.38] and 2.20 [95% CI: 1.03, 4.73; p=0.012] and fully adjusted HRs of 1.55 [95% CI: 0.50, 4.80] and 2.44 [95% CI: 1.07, 5.59; p=0.074] for moderate and hazardous/harmful drinkers, respectively), but drinking did not appear to influence mortality in the HPV-positive group, based on the fact that all of the CIs crossed one (minimally adjusted HRs of 0.81 [95% CI: 0.41, 1.59] and 0.72 [95% CI: 0.39, 1.32; p=0.301] and fully adjusted HRs of 0.85 [95% CI: 0.42, 1.71] and 0.71 [95% CI: 0.38, 1.35; p=0.402] for moderate and hazardous/harmful drinkers).

Results of the imputed analysis are presented in Appendix A4. Current smokers were at a higher risk of death compared to non-smokers, regardless of HPV status (fully adjusted HR= 2.5 [95% CI: 1.0, 6.5; *p* for trend=0.057] and 2.0 [95% CI: 1.1, 3.6; *p* for trend=0.020] for HPV-negative and HPV-positive groups, respectively). There was little evidence that alcohol consumption influenced mortality risk.

Results of the sensitivity analysis indicate that the association of smoking with survival is unlikely to be modified by HPV status (*p* for interaction in the fully adjusted model =0.092). There was weak evidence to suggest that the effect of alcohol may differ by HPV status (*p* for interaction= 0.070).

7.4.10. Interaction of tobacco and alcohol

There was little evidence of an interaction between smoking and alcohol consumption on survival in this analysis (*p* for interaction = 0.607 in the fully adjusted model).

Table 27: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by tumour site.

	All sites (n=1403)				Oral cavity (n=404)				Oropharyngeal (n=656)				Larynx (n=343)			
	HR	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**
Model 1																
Smoking																
Former	1.69	1.22	2.34		2.15	1.29	3.60		1.56	0.96	2.54		1.62	0.59	4.50	
Current	3.27	2.30	4.65	<0.001	2.13	1.19	3.83	0.012	3.97	2.33	6.76	<0.001	4.36	1.54	12.37	<0.001
Alcohol amount																
Moderate	0.74	0.53	1.05		0.68	0.38	1.22		0.84	0.47	1.48		0.72	0.38	1.35	
Hazardous/harmful	1.24	0.95	1.62	0.042	1.20	0.77	1.87	0.309	1.38	0.87	2.18	0.086	1.00	0.60	1.66	0.873
Model 2																
Smoking																
Former	1.69	1.21	2.35		2.71	1.60	4.61		1.41	0.87	2.30		1.46	0.52	4.09	
Current	2.21	1.52	3.22	<0.001	2.09	1.13	3.87	0.023	2.08	1.15	3.77	0.016	2.99	1.03	8.64	0.001
Alcohol amount																
Moderate	0.72	0.52	1.02		0.59	0.33	1.06		0.87	0.49	1.55		0.65	0.34	1.24	
Hazardous/harmful	1.08	0.83	1.41	0.330	1.04	0.66	1.62	0.673	1.18	0.75	1.87	0.351	0.98	0.59	1.64	0.856
Model 3																
Smoking																
Former	1.67	1.19	2.34		3.04	1.77	5.24		1.36	0.83	2.23		1.64	0.56	4.77	
Current	2.13	1.44	3.14	<0.001	2.22	1.16	4.25	0.009	1.66	0.88	3.12	0.068	3.45	1.13	10.53	0.001
Alcohol amount																
Moderate	0.75	0.53	1.06		0.62	0.34	1.13		0.94	0.52	1.69		0.67	0.35	1.28	
Hazardous/harmful	1.09	0.84	1.43	0.312	1.08	0.69	1.69	0.521	1.31	0.82	2.10	0.222	1.02	0.60	1.73	0.888
Model 4																
Smoking																
Former	1.63	1.16	2.28		3.15	1.82	5.45		1.29	0.78	2.14		1.60	0.55	4.66	
Current	2.03	1.37	3.01	<0.001	2.29	1.18	4.44	0.010	1.55	0.81	2.96	0.087	3.36	1.10	10.28	0.001
Alcohol amount																
Moderate	0.76	0.54	1.07		0.55	0.30	0.99		0.98	0.54	1.76		0.68	0.36	1.29	
Hazardous/harmful	1.03	0.78	1.34	0.554	0.92	0.58	1.45	0.789	1.28	0.80	2.05	0.263	0.94	0.55	1.59	0.798

*Model 1 (minimally adjusted): adjusted for age and sex; Model 2: additionally adjusted for clinical features (TNM, comorbidity, BMI, HPV); Model 3: additionally adjusted for social features (education, income, IMD and marital status); Model 4 (fully adjusted): additionally includes smoking or drinking. * test for linear trend. Values with p<0.05 are shown in bold. Abbreviations**: HR**, hazard ratio; **CI**, 95% confidence interval.*

Table 28: Association of smoking status and alcohol intake with mortality risk, stratified by tumour stage.

	Low stage (n=581)				High stage (n=822)			
	HR	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**
				Model 1				
Smoking								
Former	2.37	1.17	4.78		1.65	1.14	2.40	
Current	3.98	1.90	8.32	<0.001	3.59	2.39	5.40	<0.001
Alcohol amount								
Moderate	0.80	0.43	1.46		0.66	0.44	0.99	
Hazardous/harmful	1.44	0.90	2.32	0.080	1.10	0.80	1.52	0.268
				Model 2				
Smoking								
Former	2.20	1.09	4.44		1.44	0.99	2.10	
Current	3.00	1.42	6.35	0.004	1.99	1.28	3.08	0.002
Alcohol amount								
Moderate	0.77	0.42	1.42		0.68	0.45	1.02	
Hazardous/harmful	1.36	0.85	2.18	0.135	0.98	0.71	1.36	0.772
				Model 3				
Smoking								
Former	2.17	1.06	4.42		1.46	0.99	2.15	
Current	3.01	1.39	6.54	0.006	1.87	1.17	2.98	0.002
Alcohol amount								
Moderate	0.79	0.43	1.46		0.69	0.46	1.05	
Hazardous/harmful	1.41	0.88	2.28	0.110	0.99	0.71	1.37	0.726
				Model 4				
Smoking								
Former	2.06	1.01	4.22		1.46	0.99	2.15	
Current	2.83	1.29	6.18	0.011	1.87	1.17	2.98	0.003
Alcohol amount								
Moderate	0.78	0.42	1.44		0.73	0.48	1.11	
Hazardous/harmful	1.30	0.81	2.09	0.210	0.97	0.70	1.35	1.000

*Model 1 (minimally adjusted): adjusted for age and gender; Model 2: additionally adjusted for clinical features (TNM stage, comorbidity, BMI and HPV); Model 3: additionally adjusted for social features (education, annual household income, IMD, marital status); Model 4 (fully adjusted): additionally includes smoking or drinking. *Test for trend Abbreviations: **HR**, hazard ratio; **CI**, 95% confidence interval.*

Table 29: Association of smoking status and alcohol intake with mortality risk, stratified by HPV status.

	HPV negative (n=176)				HPV positive (n=480)			
	HR	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**
Model 1								
Smoking								
Former	5.21	1.24	21.96		1.00	0.57	1.74	
Current	6.93	1.64	29.37	0.005	2.18	0.91	5.19	0.272
Alcohol amount								
Moderate	0.82	0.28	2.38		0.81	0.41	1.59	
Hazardous/harmful	2.20	1.03	4.73	0.012	0.72	0.39	1.32	0.301
Model 2								
Smoking								
Former	4.61	1.09	19.49		1.03	0.59	1.79	
Current	3.99	0.92	17.19	0.186	2.22	0.92	5.32	0.254
Alcohol amount								
Moderate	1.10	0.37	3.28		0.83	0.42	1.65	
Hazardous/harmful	2.20	1.02	4.77	0.019	0.70	0.38	1.28	0.259
Model 3								
Smoking								
Former	5.26	1.19	23.18		1.02	0.58	1.79	
Current	3.38	0.72	15.82	0.356	2.04	0.81	5.14	0.347
Alcohol amount								
Moderate	1.72	0.56	5.32		0.82	0.41	1.65	
Hazardous/harmful	3.03	1.35	6.83	0.017	0.73	0.39	1.39	0.382
Model 4								
Smoking								
Former	3.70	0.82	16.78		1.04	0.59	1.84	
Current	2.66	0.55	12.80	0.492	2.13	0.84	5.43	0.311
Alcohol amount								
Moderate	1.55	0.50	4.80		0.85	0.42	1.71	
Hazardous/harmful	2.44	1.07	5.59	0.074	0.71	0.38	1.35	0.402

*Model 1 (minimally adjusted): adjusted for age and gender; Model 2: additionally adjusted for clinical features (TNM stage, comorbidity and BMI); Model 3: additionally adjusted for social features (education, annual household income, IMD and marital status); Model 4 (fully adjusted): additionally includes smoking or drinking. *Test for trend. Abbreviations: HR, hazard ratio; CI, 95% confidence interval.*

7.5. Discussion

7.5.1. Principal findings

This chapter assessed the potential relationships between self-reported smoking status and alcohol intake with all-cause mortality in people with cancers of the oral cavity, oropharynx and larynx. My main finding was that, even after adjusting for a wide range of prognostic factors (confounders), smoking status at the time of HNC diagnosis was associated with worse survival. In fully adjusted models, the mortality risk for current smokers was around twice that of never smokers, whilst former smokers were over 60% more likely to die during follow-up. There was no evidence of heterogeneity by treatment centre, which suggests that the relationship between smoking status and mortality risk was not influenced by factors such as the preferred treatment approach, level of care provided, or the general affluence/deprivation of the area. I found no strong evidence that alcohol intake was associated with overall mortality risk in this analysis, though participants who reported drinking moderate amounts of alcohol tended to experience better survival compared to non-drinkers.

These findings are in line with those of earlier studies, which suggest that smoking at the time of HNC diagnosis may result in poorer clinical outcomes and reduced survival [441-444]. Estimates of the size of the effect have varied considerably however, ranging from a 2.4-fold higher all-cause mortality risk in current versus never-smokers [441] to an almost fivefold higher mortality risk in people with >60 pack-years of smoking versus never-smokers [442]. There are several possible explanations for this. First, much of the evidence is based on retrospective analyses of population-level cancer registries, which are often incomplete or incorrect [619]. As a result, studies frequently have missing information on important clinical and lifestyle factors such as comorbidity, BMI, and socioeconomic position, which could potentially confound the association of smoking with survival. Those studies which have been conducted prospectively are small, typically five hundred persons or fewer [441-443], limiting their statistical power to detect the true measure of the effect. Second, estimates have been derived from different subpopulations of people, i.e. different HNC sites or tumour stages, which are often not considered separately. This could bias estimates of the effect of smoking on survival if mortality risk is greater or lesser in certain tumour groups.

With respect to alcohol consumption, the existing literature is limited and conflicting. The results of the current analysis support those of Duffy et al. who found no difference in

mortality risk between active drinkers and non-drinkers after adjusting for a wider range of confounders [441]. In contrast to this however, Mayne et al. reported a fivefold increased mortality risk for persons who drank >35 drinks per week compared to those who abstained [442]. Both previous studies were relatively small (504 and 204 people, respectively), and enrolled participants from either a single centre [441] or a single clinical trial [442], limiting their generalisability.

A J-shaped association between alcohol intake and mortality has been observed in some previous studies, though it has been suggested that the link between low-moderate intake and improved survival is not causal. Rather, higher socio-economic status is linked with moderate alcohol intake and socio-economic status is a strong predictor of longevity [620] [621].

It is biologically plausible that, as well as being risk factors for HNC, smoking could reduce survival following a diagnosis. One way in which smoking could influence survival is through its effects on treatment response. A growing body of evidence suggests that smokers have an increased risk of treatment-related adverse events and poorer clinical outcomes following radiotherapy, compared to no-smokers [622] [623]. The biological mechanisms explaining this association are not fully understood, but increased tumour hypoxia, resulting from increased carboxyhaemoglobin (the binding of carbon monoxide to haemoglobin) in smokers, is a possible explanation [624]. In addition to this, it has been suggested that tobacco reduces the efficacy of radiotherapy through triggering a p53 mutation that could promote resistance to apoptosis [625]. Smoking is also known to effect inflammatory responses [626] and immune competence [627], which could increase the likelihood of adverse clinical outcomes.

7.5.2. Strengths and limitations of the study

This study has several strengths. These include the prospective study design, the large sample size, and the ability to adjust for a wide range of biological, clinical and lifestyle covariates, including HPV. In addition to this, the risk of bias due to missing data was explored by employing a multiple imputation approach. Results of the imputed analysis was broadly consistent with those of the complete case analysis, suggesting that the presence of missing covariate data did not unduly influence the effect estimates, and hence the conclusions that can be drawn from the data.

This work does have several limitations. First, as in most previous studies, assessments of smoking and alcohol intake were based on participants' self-report; this could result in an

underestimation of the effects of smoking and drinking on HNC survival if participants under-reported their tobacco and alcohol use (see Chapter 3). Moreover, the physical characteristics of an individual's smoking behaviour, that is their "smoking topography", will influence their level of exposure to tobacco, meaning that no two "current smokers" are alike. Similarly, non-drinkers in this study make up a complex group, with some people abstaining from alcohol out of choice and some people refraining from drinking alcohol because of their health. More objective measures of tobacco and alcohol intake would provide more reliable estimates of the exposure-outcome association.

Second, although models were adjusted for multiple confounders, residual confounding by unmeasured or poorly measured factors, such as a delay in receiving treatment or other lifestyle factors related to smoking and drinking behaviours (e.g., physical activity and dietary intake), is possible.

Third, whilst the sample size was sufficient to detect the main effects of baseline smoking status and alcohol intake on survival, it was insufficient to examine interactions between these two exposures and HPV in determining mortality. It also had limited the potential to investigate whether the effects of smoking and drinking on survival were modified by cancer site. This was because there were only a small number of events (deaths) in each subgroup. The analysis was nevertheless exploratory in design, as there were no prior hypotheses that smoking or alcohol intake would have a greater or lesser effect on survival in any one cancer group.

Fourth, HPV status was determined based on HPV-specific antibody levels but, as discussed in Chapter 5, the presence of HPV16 RNA is considered the gold standard measure. That said, previous studies have confirmed that detection of antibodies against E6 and E7 oncoproteins shows good correlation with HPV DNA in the tumour tissue. In their study, Kreimer *et al*. showed that high HPV viral load increased the odds of HPV16 E6 seropositivity 57-fold and HPV16 E7 seropositivity 26-fold [628]. In addition to this, HPV16 E6 antibodies have also been shown to be independent favourable prognostic factors in OPCs [551 552]. Some HPV-related OPCs are mediated by other HPV genotypes, including HPV18 (1–8% of OPCs), and less commonly HPV33, −35, −56 and −67 [629]. Only HPV-16 was considered in the current analysis. This was because 73% of the participants with OPCs were HPV16 E6 positive, compared to just 1% who were HPV18 E6 positive.

Finally, it was not possible to examine whether baseline smoking status and alcohol intake influenced cancer-specific mortality as cause-specific mortality data were not available for all

participants at the time of this analysis. Previous studies suggest that death from non-cancerous causes (competing mortality) and second primary malignancies are important events in HNC [630] and could provide greater insight into the biological mechanisms that underlie the associations of smoking and drinking with survival. It is important to note however, that cause of death information death certificates is often inaccurate [631]. Accuracy of all-cause mortality is solely dependent on the number of deaths identified and is arguably a more reliable outcome measure.

7.5.3. Conclusions

This chapter shows that smoking status at the time of HNC diagnosis is associated with poorer survival and may improve prognostication in this population, but there was no apparent difference in mortality risk among people who consumed different amounts of alcohol pre-diagnosis. The confidence in the model estimates is curbed by the fact that exposure levels were based on participants' self-report. DNA-methylation based biomarkers of tobacco and alcohol consumption can provide more robust estimates of exposure than self-reported intake, as discussed in Chapter 3. With that knowledge in mind, the ensuing chapter looks to investigate the relationships between DNA methylation predictors of smoking and alcohol (and other lifestyle traits) and mortality risk in H&N5000.

Chapter 8: Epigenetic prediction of complex traits and mortality in oropharyngeal cancer in H&N5000

This chapter includes sections from the publication below:

Langdon R.J*., Beynon R.A*., Ingarfield K., Marioni R.E., MacCartney D.M., Martin R.A., Ness R.A., Pawlita M., Waterboer T., Relton C., Thomas S.J., Richmond R.C. Epigenetic prediction of complex traits and mortality in a cohort of individuals with oropharyngeal cancer (2020) *Clinical Epidemiology*. 12

* These authors contributed to the manuscript equally.

8.1. Introduction

The findings of the previous chapter confirm that smoking status at the time of HNC diagnosis is associated with survival, but there was no strong evidence to show that alcohol intake influenced mortality outcomes. Whilst this analysis was one of the largest of its kind to date, estimates of the effects of these modifiable behaviours on survival were based on participant's self-report which, for reasons that have already been discussed at length in this thesis, are often unreliable. Moreover, smoking and alcohol exposures were pigeon-holed into three categories – never-, former-, or current smoker and non-drinker, moderate-drinker, or hazardous-to-harmful drinker. In reality, these behaviours do not sit neatly into groups but represent a continuum of exposure- a cumulative 'hit' [632]. Consequently, the results of this analysis may not fully capture the relevant smoking and alcohol-related behaviour and its relationship with mortality risk. Such limitations underline the need for objective measures of lifestyle exposures for precise classification in epidemiological studies.

In Chapter 3, I described how biological predictors of certain lifestyle traits can improve exposure assessment and may facilitate improved risk prediction in epidemiological and public health settings. Advances in epigenome-wide profiling have permitted the identification of "epigenetic signatures" (i.e. DNAm patterns), for a variety of lifestyle factors associated with health and mortality. These include DNAm predictors for smoking, alcohol intake, BMI, waist-to-hip ratio, body fat percentage, HDL cholesterol and education [509 511 633]. These predictors explain varying proportions of the phenotypic variance seen in their

respective traits, with methylation-based predictors of smoking demonstrating particularly good discrimination between current and never smokers (AUC=0.98) [633].

The relationship between DNAm scores and all-cause mortality have been assessed among general populations, but there is currently a paucity of studies conducted in clinical settings. In one study, which used data from the Lothian Birth Cohort 1936 (LBC1936), higher mortality risk was associated with higher DNAm scores for smoking (HR = 1.29, 95% CI: 1.05, 1.57; *p* = 0.013), alcohol consumption (HR = 1.24, 95% CI: 1.08, 1.43; *p* = 0.003), and waist-to-hip ratio (HR = 1.24, 95% CI: 1.08, 1.42; *p* = 0.002), whilst a higher DNAm score for education was associated with lower mortality risk (HR = 0.81, 95% CI = 0.71, 0.93; *p* = 0.004) [633]. In this chapter, I report on a novel application of DNAm risk scores for health and lifestyle related traits to a clinical cohort of individuals with head and neck cancer - H&N5000.

8.2. Aims and objectives

The specific aims of this analysis were, firstly, to assess whether externally derived DNAm predictors for 4 complex traits, namely smoking, alcohol consumption, BMI and educational attainment, could provide an accurate prediction of directly-measured phenotypes in a subset of participants with oropharyngeal tumours (OPC) and secondly, to examine the association of these DNAm predictors with mortality risk in the same sub-set of individuals with OPC, using Cox proportional hazards regression models.

8.3. Methods

8.3.1. Study population

The study population for this analysis included a sub-set of H&N5000 participants with OPC (selected based on ICD-10 coding- pathological where available), for whom OncoChip genotype data, baseline question and data capture data were available (n=448). Full details of the study method and overall population from which these participants were drawn are described in Chapter 4.

8.3.2. Assessment of tobacco, alcohol, BMI, and education

Information on participants' smoking histories, use of alcohol prior to receiving their HNC diagnosis, highest educational attainment and BMI were obtained at baseline via the self-

administered questionnaires, as described in chapter 4. In line with the previous analysis, smoking categories were preserved as “current”, “former” and “never” and drinking categories were grouped as “non-drinker”, “moderate-drinker” and “hazardous to harmful-drinker”, based on current UK guidelines (i.e. moderate drinkers includes those drinking <14 units of alcohol a week and hazardous to harmful-drinkers includes those drinking 14 units or more each week). BMI was calculated as weight (kg) per height (m) squared. Education levels were defined as “school”, “college” (which includes sixth form and further education), or “Degree” (which includes polytechnic or university).

8.3.3. Epigenetic profiling and pre-processing

The epigenetic data for this analysis were made available thanks in large part to work by Matthew Suderman, Rebecca Richmond, and Ryan Langdon at the Integrative Epidemiology Unit (IEU), University of Bristol. Full details are provided in the “Methods for collecting and processing materials” section of Chapter 5. Briefly, following extraction, DNA was bisulphite-converted using the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA, USA). Genome-wide methylation status of over 850,000 CpG sites was then measured using the Infinium MethylationEPIC BeadChips (EPIC array) (Illumina, USA), according to the manufacturer’s protocol. The arrays were scanned using an Illumina iScan (version 2.3).
Raw data files (IDAT files) were pre-processed and normalised using the R package *meffil* (https://github.com/perishky/meffil/) [634], which excluded 8 of the total samples for not meeting the quality control criteria (n=440/448). A further 32 samples were removed following pathological re-classicisation (i.e. the tumours from which the samples originated did not originate in the oropharynx), leaving a final sample of n=408 (Figure 43).

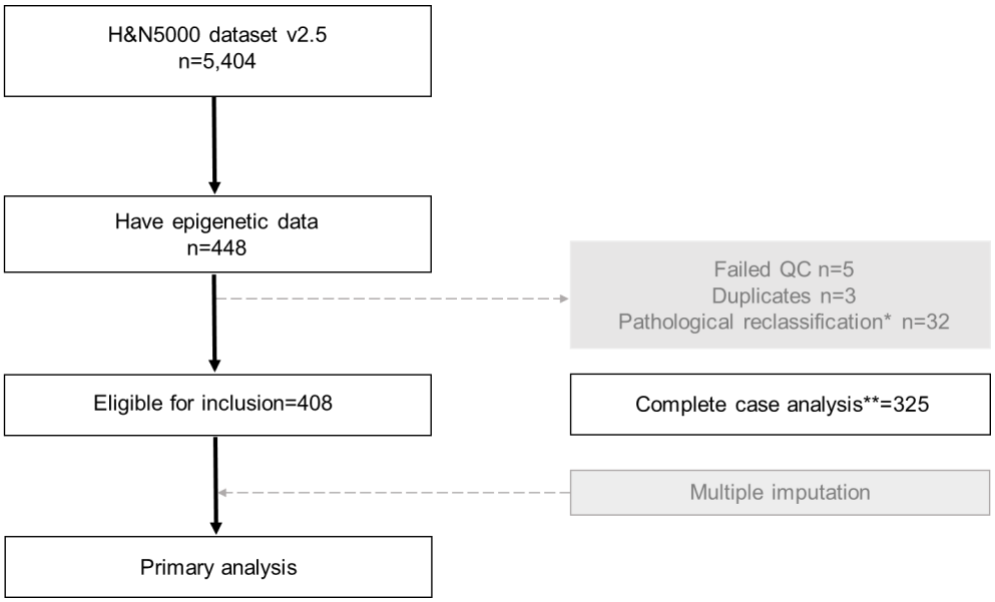
8.3.4. Epigenetic risk score generation

DNAm predictors for smoking, alcohol, education, and BMI were computed for each participant using different sets of coefficients made available from the largest or most recently published EWAS of methylation of CpG sites in relation to these traits. Scores were calculated as $\beta_1 M_1 + \beta_2 M_2 + \ldots + \beta_n M_n$, where β refers to the β-regression coefficient of the selected CpG and M refers to the normalised methylation level at that locus (in H&N500). An overview of the different scores, including the number of CpG sites included in each model, is provided in Table 30. DNA methylation at cg05575921 (*AHRR*) was also determined for each participant.

8.3.5. Multiple imputation

The number of participants with OPC and epigenetic data available is relatively small. To avoid case deletion due to missing data on baseline covariates, primarily BMI which was missing for 33% of participants (Table 31), missing data were imputed by multiple imputations. As mentioned in the previous chapter, information on participants height and weight was not collected at the start of the study, meaning that BMI could not be calculated for all participants. The missingness mechanism was presumed to be MAR because it could be explained by other variables in the dataset (i.e. date of enrolment). Missing values were imputed using the ICE package for multiple chained equations in STATA, as per the previous chapter. Twenty imputed datasets were generated and then combined using Rubin's rules [614]. The imputation model included the event indicator, the Nelson-Aalen estimator of the cumulative hazard, all the variables that were used in substantive Cox models and any other available variables that could help to explain the missing data, such as household income and marital status. .

Figure 43: Flow of participants included in the analysis.



** OPC cases were selected for epigenetic analysis using clinical ICD codes. v2.5 of the HN5000 dataset uses Pathology ICD codes where available.** Data available for age, gender, TNM stage, HPV status, comorbidity, education, BMI, self-reported smoking status and alcohol consumption.*

Table 30: An overview of the DNAm scores employed in the current analysis.

DNAm predictor	Model development	CpG sites	Ref.
Smoking status			
Joehanes Bonferroni	Linear mixed models (also called "multilevel models") were conducted, then combined in a random-effects model meta-analysis. One set of CpGs was selected based on a Bonferroni threshold of $P<1 \times 10^{-7}$ (485,381 tests) and another was selected based on a genome-wide FDR threshold of $P<0.05$.	2623	509
Joehanes FDR		18760	
Zhang	An epigenome-wide screen for cotinine associated loci was first carried out in a discovery set (n=500) using median quantile regression and a Bonferroni threshold $p<1.13$ x1 0^{-7}. 61 CpGs were taken forward to the replication phase (n=500). 40/61 CpGs replicated in the validation set ($p<8.2 \times 10^{-4}$). The final DNAm score was developed through stepwise selection (with significance levels for deletion from and adding to the model of 0.01) of the confirmed CpGs.	3	635
McCartney smoking	The DNAm predictors were built on a subset of 5087 individuals from GS using penalized regression models. Tenfold cross-validation was applied and the mixing parameter (alpha) was set to 1 to apply a LASSO penalty. Coefficients for the model with the lambda value corresponding to the minimum mean cross-validated error were extracted and applied to an independent validation cohort, the LBC (n=895), to create the DNAm predictors.	233	633

Table 30 continued.

DNAm predictor	Model development	CpG sites	Ref.
Alcohol consumption			
McCartney alcohol	See McCartney smoking.	450	633
Liu model 1	First, a meta-analysis was performed in 8 European ancestry cohorts using an inverse-variance weighted random-effects model and CpGs were selected at a threshold $P<5 \times 10^{-6}$. CpGs selected from the meta-analysis were then taken forward and simultaneously included in a LASSO regression, which was performed in an independent cohort (the FHS cohort). Four sets of CpGs were selected using s=0.08 (model 1), s=0.12 (model 2), s='lambda.1se' (model 3) and s='lmabda.min' (model 4), which represents the penalty apllied. The criterion s= 'lambda.min' selected the largest number of CpGs and s=0.12 yielded the most parsimonious set of CpGs.	5	511
Liu model 2		23	
Liu model 3		78	
Liu model 4		144	
Body mass index			
McCartney BMI	See McCartney smoking.	1109	633
Educational attainment			
McCartney education	See McCartney smoking.	373	633

Abbreviations: ***CpG****, cytosine-phosphate-guanine site;* ***FDR****, false-discovery rate;* ***FHS****, Framingham Heart Study;* ***GS****, Generation Scotland;* ***LASSO****, least absolute shrinkage elector operator;* ***LBC****, Lothian birth cohort;* ***Ref****., reference.*

Table 31: Proportion of missing data in the epigenetic dataset (n=408).

Variable	% Missing	% Missing
Age	5	1.23
Gender	0	0
Ethnicity	10	2.45
TNM	0	0
Treatment group	0	0
HPV status	0	0
Comorbidity	3	0.74
BMI	136	33.33
Annual household income	51	12.5
Education	19	4.66
Marital status	6	1.47
Smoking status	16	3.92
Alcohol consumption	8	1.96

Statistical analysis

8.3.5.1. Associations of epigenetic scores with self-reported phenotypes

Linear regression analyses of epigenetic risk scores against directly measured phenotypes (i.e. self-reported traits) were performed to determine how much of the variance in each phenotype was explained by the corresponding score(s). This was undertaken in two-stages: in the first stage, the DNAm score was regressed on age and gender; in the second stage, the directly measured phenotype was regressed on the residual obtained from stage 1. The percentage of the variance explained was captured in the R_2 statistic from the second model.

8.3.5.2. Proportion of variance in survival explained by DNAm scores

The variation in the outcome (survival) accounted for through the DNAm predictors was assessed using an analogue of the R_2 for censored survival data [616] and implemented using the `str2d` command in Stata.

8.3.5.3. Survival analysis

The end point of this study was all-cause mortality, defined as the time in days from study enrolment to date of death from any cause or the date of censorship (i.e., the last date of follow-up). Kaplan-Meier survival curves and log-rank test were used to investigate the univariate impact of covariates on mortality (Appendix B1). Mortality risk was then assessed

in relation to each of the DNAm scores using Cox proportional-hazards models. Given that the DNAm scores have different scales (Appendix B2-B13), I standardise them to facilitate direct comparison between scores. Using smoking as an example, I standardised the methylation scores of individuals x_i by dividing by the difference in mean DNAm scores between current and never smokers m_c-m_n as follows:

$$\text{Standardised score} = x_i / (m_c - m_n)$$

Standardising each of the methylation scores in this way means that the HR per unit increase in standardised score is comparable across all the methylation scores. The interpretation of the HR is:

HR = the change in risk per unit of standardised methylation score, where 1 unit increase in standardised methylation score corresponds to the average increase in the measured methylation score experienced by current smokers compared with never smokers.

.

An example of the code used to standardise DNAm scores is provided in the supporting material (Appendix B14). For alcohol, BMI, and education, using the same equation, I calculated the average increase in score experienced by non-drinkers compared with hazardous to harmful drinkers, overweight people compared with not overweight people, and degree-educated people compared with school-educated people. Individuals were classed as being overweight if they had a BMI >25 and not overweight if they had a BMI of ≤24. NHS cut-offs define someone who is healthy as having a BMI of 18-25 and someone who is underweight as having a BMI of <18 but since only 4/408 people fell into the latter (range 15.6 to 17.9), it was decided that these groups would be combined into a single “not overweight” category.

Four separate models were fit to examine the relationship between DNAm scores with mortality:

1) a minimally adjusted model that adjusted for age, gender, cell counts and batch effects;
2) a ‘clinical model’ that additionally adjusted for tumour stage, HPV status and comorbidity (i.e. information that would be available to physicians);

3) a model that additionally adjusted for the corresponding directly measured phenotype (e.g. models that examined the association of smoking related DNAm scores with mortality adjusted for self-reported smoking status);
4) a model that additionally adjusted for the other directly measured phenotypes.

Cell types included B-cells, CD4+ T-cells, CD8+ T-cells, eosinophils, monocytes, neutrophils and natural killer cells. Counts were estimated in whole blood using the method of Houseman *et al* [636]. Clinical variables were selected based on prior knowledge of established risk factors (see Chapter 2). Results of the imputed analysis are presented here. A complete case adjusted analysis was also performed as a sensitivity analysis.

The PH assumption was tested for each of the covariates included in the Cox model (without adjustment) using a combination of graphical and statistical tests based on Schoenfeld residuals, using the `estat phtest` command in stata (which tests for a non-zero slope of Schoenfeld residuals versus time). Covariate-specific and global tests were then performed after fitting the adjusted survival models to ensure that the assumption held. Graphical and formal tests indicated that the variables HPV status and comorbidity may not satisfy the PH assumption (see Appendix B15 and B17) for Kaplan–Meier observed survival curves versus Cox predicted curves and B16 and B18 for log-log survival curves). However, the statistical power to detect violations is limited due to the limited number of events (i.e. deaths) in the current analysis [637], particularly after 3-years (see Appendix B1) As a further sensitivity analysis, the data were split (i.e. censored) at 3-years and the above analyses were repeated.

8.4. Results

8.4.1. Baseline characteristics of the study population

The study sample was described in detail earlier in Chapter 6 (Tables 21 and 22). Overall, 78% of people included in the analysis were male and 70% had HPV-positive tumours. There were differences in clinical and lifestyle characteristics between HPV-positive and HPV-negative individuals, including differences in tumour stage (HPV-positive individuals tended to be diagnosed at a later tumour stage), income (HPV-negative individuals tended to earn less as a household) and smoking status (HPV-negative individuals were more likely to be current smokers).

Histograms showing the distribution of DNAm scores stratified by their respective directly-measured phenotypes are presented in Appendices B2 to B13. AHRR was unique in

displaying a bimodal distribution, with higher DNAm scores in non-smokers. Mean DNAm scores stratified by 3-year mortality are provided below in Table 32. There was strong evidence of a difference between groups in all of the scores except for the education and BMI-related DNAm scores derived by McCartney.

The baseline descriptives of the analytic dataset (i.e. individuals with OPC and epigenetic data available) were compared against participants without epigenetic data available (all other people with OPC in HN5000). Overall, the sample of people with epigenetic data were fairly representative of the OPC population as a whole (Table 33), with the exception that there was evidence of a difference in comorbidity levels between groups (p=0.034). In the group without DNAm data, 46% had no comorbidity, 36% had mild comorbidity and 18% had severe comorbidity. This compares to 52%, 29% and 19%, respectively, in the group with DNAm data available.

Table 32: Average DNAm values for people who are alive at 3-years compared to people who are dead at 3-years.

	Dead at 3 years (n=77)	**Alive at 3 years (n=331)**	
DNAm Score	*mean (SD)*	*mean (SD)*	*p-value*
	Smoking		
AHRR	0.06 (0.22)	0.14 (0.97)	<0.001
Joehanes Bonferroni	0.03 (0.13)	-0.04 (0.25)	0.001
Joehanes FDR	-0.01 (0.08)	-0.02 (0.19)	<0.001
Zhang	-0.02 (0.11)	-0.02 (0.50)	0.038
McCartney	-0.47 (1.01)	-0.02 (0.21)	0.002
	Alcohol consumption		
McCartney	0.06 (0.22)	-0.01 (0.12)	0.013
Liu 5 CpG	0.08 (0.25)	-0.02 (0.23)	0.001
Liu 23 CpG	0.03 (0.14)	-0.01 (0.14)	0.030
Liu 78 CpG	0.18 (0.54)	-0.05 (0.49)	<0.001
Liu 144 CpG	0.29 (0.84)	-0.07 (0.74)	<0.001
	Educational attainment		
McCartney	0.12 (0.29)	0.00 (0.07)	0.127
	BMI		
McCartney	0.12 (0.67)	0.00 (0.11)	0.103

*Abbreviation: **SD**, standard deviation.*

Table 33: A comparison of the baseline characteristics of people with epigenetic data and people without epigenetic data available (OPC only).

	Without epigenetic data (n=1,178)		**With epigenetic data (n=408)**		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*p-value*
Gender					
Male	1178	79.20%	317	77.70%	
Female	310	20.80%	91	22.30%	0.519
TNM stage					
I	66	4.50%	17	4.20%	
II	145	9.80%	39	9.60%	
III	210	14.20%	58	14.20%	
IV	1059	71.60%	294	72.10%	0.993
HPV serology group					
Negative	384	32.80%	122	29.90%	
Positive	788	67.20%	286	70.10%	0.286
Comorbidity					
None	665	45.70%	211	52.10%	
Mild	524	36.00%	119	29.40%	
Moderate/Severe	267	18.30%	75	18.50%	0.034
Education level					
School education	422	44.10%	170	43.70%	
College	340	35.50%	158	40.60%	
Degree	196	20.50%	61	15.70%	0.071
Annual household income					
<£18,000	340	38.10%	138	38.70%	
£18000-£34,999	277	31.10%	103	28.90%	
>£35,000	275	30.80%	116	32.50%	0.722
IMD					
Low Deprivation	531	39.30%	149	38.80%	
Moderate Deprivation	306	22.60%	81	21.10%	
High Deprivation	515	38.10%	154	40.10%	0.723
Relationship status					
Single (never married)	109	11.00%	47	11.70%	
Currently in relationship	704	71.00%	280	69.70%	
No longer with spouse	178	18.00%	75	18.70%	0.871
Smoking status					
Never	252	26.50%	110	28.10%	
Former	542	56.90%	205	52.30%	
Current	158	16.60%	77	19.60%	0.246
Alcohol consumption					
Non-drinker	229	23.60%	104	26.00%	
Moderate drinker	216	22.20%	90	22.50%	
Hazardous-harmful drinker	527	54.20%	206	51.50%	0.578
	N	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	1479	59.19 (9.06)	403	58.43 (9.63)	0.145
BMI	826	26.83 (5.24)	272	26.45 (4.94)	0.283

8.4.2. Correlation between covariates

The Pearson's correlation coefficient's (r) for each pair of covariates are presented in Table 34. There was strongest evidence of an association between the following covariates: HPV status and smoking (r=-0.36; p<0.001), age and comorbidity (r=0.34; *p*<0.001), gender and alcohol intake (r=-0.25; p<0.001), BMI and smoking (r=0.22-; p<0.001), HPV status and comorbidity (r=-0.19; p<0.001), and HPV status and TNM stage (r=0.18; *p*<0.001). There were further associations between HPV status and BMI (r=0.19; p<0.01) comorbidity and BMI (r=0.15; p<0.05) and education and smoking (r=-0.13; p<0.05) and age and HPV status (r =-0.16; p<0.1).

8.4.3. Correlation between DNAm predictors

As expected, the two smoking-related DNAm scores developed by Joehanes were strongly correlated with one another (r = 0.92; *p*<0.001), as were the alcohol DNAm scores developed by Lui (values ranging from r= 0.65; *p*<0.001 to r= 0.96; *p*<0.001) (Table 35). There was a strong positive correlation between the Joehanes smoking predictors and both the AHHR DNAm score (r=0.63 and 0.82; *p*<0.001 for Joehanes FDR and Joehanes Bonferroni, respectively) and the smoking DNAm score developed by McCartney (r=0.82; *p*<0.001). There was strong evidence of a moderate negative correlation between the DNAm predictor for education and the DNAm predictors for smoking (r values of between -0.17; *p*<0.001 to -0.30; *p*<0.001). The BMI DNAm score was negatively correlated with the AHRR DNAm score (r=-0.12; *p*<0.05) and the alcohol DNAm score (r=-0.22; *p*<0.001). The BMI predictor was not correlated with any of the smoking scores in this analysis.

8.4.4. Proportion of variance in phenotype explained by the DNAm scores

Age and gender-adjusted linear regression models showed that the DNAm predictors explained a small proportion of the phenotypic variance in educational attainment (0.7%), a moderate proportion of the variance in BMI (22%) and alcohol (7-16%) and a variable proportion of the variance in smoking (5-49%) (Table 36). The DNAm predictor that explained the greatest proportion of the variance in alcohol consumption was the DNAm score developed by Liu *et al* based-on methylation at 144 CpG sites. The methylation status of AHRR explained the greatest proportion of the variance in smoking (49%), followed by the McCartney score (44%). The Zhang DNAm predictor explained the least amount of the variance in smoking (5%).

Table 34: Pearson's correlation coefficient matrix for included covariates.

	Age	Gender	TNM stage	HPV status	Comorbidity	BMI	Education	Smoking	Alcohol
Age	1								
Gender	-0.02 (-0.12, 0.08)	1							
TNM stage	-0.03 (-0.12, 0.07)	0.03 (-0.07, 0.12)	1						
HPV status	-0.16** (-0.25, -0.06)	0.02 (-0.08, 0.11)	0.18*** (0.08. 0.27)	1					
Comorbidity	0.34*** (0.26, 0.42)	-0.09 (-0.18,0.01)	-0.06 (-0.16, 0.03)	-0.19*** (-0.28, -0.09)	1				
BMI	-0.07 (-0.19, 0.05)	-0.10 (-0.22, 0.02)	0.03 (-0.09, 0.15)	0.19** (0.08, 0.31)	0.15* (0.03, 0.26)	1			
Education	-0.10 (-0.20, 0.00)	-0.05 (0.14, 0.05)	-0.01 (-0.11, 0.09)	0.02 (-0.08, 0.12)	-0.06 (-0.16, 0.04)	-0.03 (-0.15, 0.09)	1		
Smoking	0.05 (-0.06, 0.14)	-0.05 (-0.15,0.05)	-0.05 (-0.15, 0.05)	-0.36 (-0.44, -0.27)	0.18*** (0.09, 0.28)	-0.22*** (-0.33, -0.10)	-0.13* (-0.23, -0.03)	1	
Alcohol intake	-0.06 (-0.16, 0.04)	-0.25*** (-0.34, -0.16)	-0.03 (-0.34, -0.16)	-0.08 (-0.17, 0.02)	0.03 (-0.7,0.13)	0.02 (-0.10, 0.14)	0.09 (-0.02, 0.18)	0.10 (-0.00, 0.20)	1
	p<0.05*		p<0.01**		p<0.001***				

*Abbreviations: **BMI**, body mass index; **HPV**, human papillomavirus; **TNM**, tumour-node-metastasis. Missing values were handled by pairwise deletion, meaning all available observations are used to calculate each pairwise correlation (as opposed to listwise deletion in which the entire observation is omitted from the estimation sample if any of the variables are missing for that observation).*

Table 35: Pearson's correlation coefficient matrix for DNAm scores.

	AHRR	Joe FDR	Joe Bonferroni	Zhang	Smoking	Alcohol	BMI	Education	Liu 5 CpG	Liu 23 CpG	Liu 78 CpG	Liu 144 CpG
AHRR	1											
Joe FDR	0.63*** (0.69, 0.57)	1										
Joe Bonferroni	0.82*** (0.85, 0.78)	0.92*** (0.91, 0.94)	1									
Zhang	0.29*** (0.37, 0.19)	0.02 (0.08, 0.12)	0.10* (0.01, 0.20)	1								
Smoking	0.82*** (0.85, 0.78)	0.53*** (0.45, 0.56)	0.68*** (0.62, 0.73)	0.32*** (0.23, 0.40)	1							
Alcohol	0.37*** (-0.45, -0.28)	0.36*** (0.28, 0.44)	0.35*** (0.26, 0.43)	0.08 (-0.02, 0.17)	0.22*** (0.12, 0.30)	1						
BMI	-0.12* (-0.02, -0.21)	-0.10 (-0.19, 0.44)	-0.10 (-0.19, 0.01)	0.00 (-0.10, 0.10)	-0.07 (-0.17, 0.02)	-0.22*** (-0.31, -0.12)	1					
Education	-0.25*** (-0.16, -0.34)	-0.17*** (-0.26, -0.8)	-0.27*** (-0.36, -0.18)	-0.19*** (-0.28, -0.10)	-0.30*** (-0.39, -0.21)	0.07 (-0.03, -0.12)	-0.11* (-0.20, -0.01)	1				
Liu 5CpG	0.27*** (0.36, 0.18)	0.33*** (0.24, 0.41)	0.43*** (0.35, 0.51)	0.05 (-0.05, 0.15)	0.19*** (0.10, 0.29)	0.31*** (0.22, 0.40)	0.05 (-0.05, 0.14)	-0.28*** (-0.36, -0.18)	1			
Liu 23 CpG	0.18*** (0.28, 0.09)	0.24*** (0.15, 0.33)	0.34*** (0.25, 0.42)	0.05 (-0.05, 0.14	0.14** (0.05, 0.24)	0.26*** (0.16, 0.34)	0.04 (-0.06, 0.13)	-0.29*** (-0.38, -0.20)	0.95*** (0.94, 0.96)	1		
Liu 78 CpG	0.23*** (0.32, 0.14)	0.33*** (0.24, 0.41)	0.38*** (0.30, 0.46)	0.02 (-0.8, 0.12)	0.14** (0.05, 24)	0.43*** (0.34, 0.50)	0.05 (-0.05, 0.15)	-0.21*** (-0.30, -0.12)	0.88*** (0.85, 0.90)	0.77*** (0.73, 0.81)	1	
Liu 144 CpG	0.21*** (0.30, 0.11)	0.28*** (0.19, 0.37)	0.32*** (0.23, 0.40)	0.04 (-0.06, 0.13)	0.12* (0.02, 0.21)	0.47*** (0.39, 0.53)	0.06 (-0.03, 0.16)	-0.15** (-0.24, -0.05)	0.76*** (0.71, 0.80)	0.65*** (0.59, 0.70)	0.96*** (0.95, 0.97)	1

p<0.05* p<0.01** p<0.001***

*Abbreviations: **Joe Bonferroni**, Joehanes Bonferroni-adjusted smoking scores; **Joe FDR**, Joehanes false discovery rate-adjusted smoking scores; **Smoking**, McCartney smoking scores; **Alcohol**, McCartney alcohol scores; **Education**, McCartney education scores; **BMI**, McCartney body mass index scores; **Liu 5**, Liu alcohol scores based on a set of 5 CpGs; **Liu 23**, Liu alcohol scores based on a set of 23 CpGs (the base cytosine (C) linked by a phosphate bond to the base guanine (G) in the DNA nucleotide sequence); **Liu 78**, Liu alcohol scores based on a set of 78 CpGs; **Liu 144**, Liu alcohol scores based on a set of 144 CpGs. Missing values were handled by pairwise deletion.*

Table 36: The proportion of phenotypic variance explained in each trait by the respective DNAm-based predictor.

DNAm Score	Variance explained (%)
Self-reported smoking	
McCartney	44.14
AHRR	48.65
Joehanes (FDR)	22.10
Joehanes (Bonferroni)	38.41
Zhang	5.05
Self-reported alcohol	
McCartney	7.19
Liu 5 CpG	14.42
Liu 23 CpG	10.90
Liu 78 CpG	15.70
Liu 144 CpG	16.00
BMI	
McCartney	21.53
Education level	
McCartney	0.68

8.4.6. Proportion of variance in survival explained by the DNAm scores

Table 37 compares the explained variation in survival for models that included age, gender, cell counts and batch effects and each of the twelve DNAm-based predictors, respectively. The basic model that included age and gender accounted for 24% of the variation in survival. The addition of DNAm predictors for smoking added between 6% (Joehanes FDR R_2=0.30 [95% CI=0.20,0.40]) and 10% (AHRR R_2=0.34 [95% CI=0.23, 0.44]) to the proportion of variance explained. Models that included DNAm predictors for alcohol intake, BMI and educational attainment did not explain much more of the variation in survival than the basic model including only age and gender (Table 37).

Table 37: The proportion of variance in survival explained by DNAm-based predictors for smoking, alcohol intake, BMI and educational attainmetn (n=408).

Model	R^2	95% Confidence interval Lower	Upper
Basic model*	0.24	0.14	0.35
Additionally adjusted for DNAm predictors:			
Smoking			
AHRR	0.34	0.23	0.44
McCartney	0.31	0.20	0.41
Joehanes FDR	0.30	0.20	0.40
Joehanes Bonferroni	0.32	0.21	0.41
Zhang	0.29	0.19	0.40
Alcohol			
McCartney	0.26	0.16	0.36
Lui 5 CpG	0.27	0.17	0.37
Lui 23 CpG	0.25	0.15	0.35
Lui 78 CpG	0.26	0.16	0.36
Lui 144 CpG	0.26	0.16	0.36
BMI			
McCartney	0.26	0.16	0.37
Educational attainment			
McCartney	0.26	0.16	0.37

*Includes age, gender, cell counts (CD4 T-cells, CD8 -cells, eosinophils, monocytes neutrophils natural killer) and batch effects.

8.4.7. Association of DNAm predictors with all-cause mortality

There were 105 deaths during a median follow-up time of 5.4 years (Inter quartile range (IQR)=4.9 to 6.0 years). The results of the Cox regression analyses are presented in Figures 44 a. to 44 d. In minimally adjusted models, which controlled for age, gender, cell counts and batch effects, all of the smoking and alcohol DNAm predictors were strongly associated with mortality risk (Figure 44 a.), with HRs ranging from 1.28 (95% CI: 1.12, 1.47; p=4.07 x 10^{-4}) to 2.99 (95% CI: 1.96, 4.56; p=3.67 x 10^{-7}) for smoking and 1.15 (95% CI: 1.02, 1.28; p=1.69 x 10^{-2}) to 1.19 (95% CI: 1.07, 1.32; p=1.78 x 10^{-3}) for alcohol. There was evidence that higher BMI and higher educational attainment scores were protective (HRs= 0.88, 95% CI: 0.77, 1.01; p=6.25 x 10^{-2} and 0.94, CI: 0.89, 1.00; p=4.57 x 10^{-2}, respectively). When the models were additionally adjusted for clinical factors (tumour stage, HPV status and comorbidity), there was little evidence that the DNAm predictors for education or the DNAm

score for alcohol (based on a set of 23 CpGs) (Figure 44 b.) were associated with survival. On further adjusting for the corresponding directly measured phenotypes, only AHHR, Joehanes and Zhang DNAm scores for smoking were associated with mortality risk (HR= 1.90, 95% CI: 1.06, 3.38; p=3.05x10^{-2} for AHRR, 1.56 (95% CI: 1.04, 2.34; p=3.12x10^{-2} for Joehanes (FDR), and 1.90 (95% CI: 1.17, 3.09; p=9.6 x 10^{-3} for Joehanes (Bonferroni); Figure 44 c.). Adjusting for the other directly measured phenotypes (fully adjusted model) did not alter the effect estimates (Figure 44 d.).

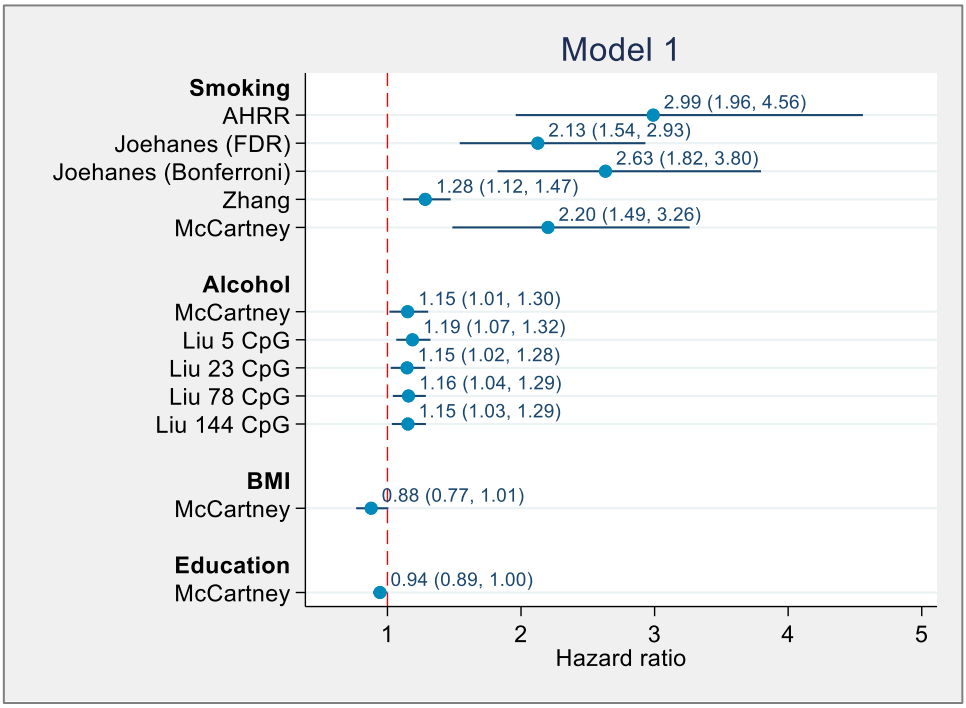
In the first sensitivity analysis, which included 248 people with complete data, the Joehanes scores and the Zhang score were also associated with mortality risk in the fully adjusted model. The effect estimate for the Zhang score was broadly comparable (1.19, 95% CI: 1.0, 1.42; p=5.02 x 10^{-2}) but HRs were higher for the Joehanes scores (HR=1.95, 95% CI: 1.11, 3.45; p=2.08 x 10^{-2} for the score derived using an FDR significance threshold and HR = 2.11, 95% CI: 1.08, 4.12; p=2.99 x 10-2 for the score derived using a strict Bonferroni threshold). The AHRR DNAm score was not associated with survival following adjustment for self-reported smoking in the complete case analysis. In comparison to the primary analysis, three of the alcohol predictors developed by Liu *et al*, based on 5, 78 and 144 alcohol related CpGs, were associated with survival, with HRs of between 1.26 (95% CI: 1.03, 1.54; p=2.42 x 10^{-2}) and 1.29 (95% CI: 1.05, 1.58; p=1.60 x 10^{-2}). A comparison of the Cox results for the primary analysis (n=408) and the complete case analysis (n=248) can be found in Appendix B19.

When the data were censored at 3 years (77 deaths), the effect estimates for the association of smoking related DNAm scores with survival were higher than in the primary analysis, but the CIs were wider (Appendix B18 to B21). In the minimally adjusted model, all of the smoking scores except the DNAm score developed by Zhang were associated with survival (AHRR HR= 3.63 [95% CI: 1.88, 6.99; p=1.20 x 10^{-4}]; Joehanes FDR HR= 2.58 [95% CI: 1.57, 4.23; p=1.76 x 10^{-4}]; Joehanes Bonferroni HR=3.36 [95% CI: 1.91, 5.91; p=2.57 x 10]; McCartney HR= 3.31 [95% CI: 1.80, 6.10; p=1.21 x 10^{-4}]). There was weak evidence to suggest an association between DNAm-predictors of alcohol and mortality risk, with HRs ranging from 1.10 (95% CI: 0.91, 1.34) to 1.20 (95% CI: 1.02, 1.42). Neither BMI nor education predictors were related to survival. The association of smoking-related DNAm scores with survival that were observed in the minimally adjusted model remained but attenuated after controlling for clinical variables and directly measured phenotypes (AHRR HR= 2.73 [95% CI: 1.08, 6.89; p=3.37 x 10^{-2}]; Joehanes FDR HR=2.09 [95% CI: 1.10, 3.95 p=2.36 x 10^{-2}]; Joehanes Bonferroni HR= 2.85 [95% CI: 1.33, 6.13 p=7.24 x 10^{-3}]; McCartney HR= 2.50 [95% CI: 1.10, 5.69; p=2.92 x 10^{-2}]), but there was no apparent relationship

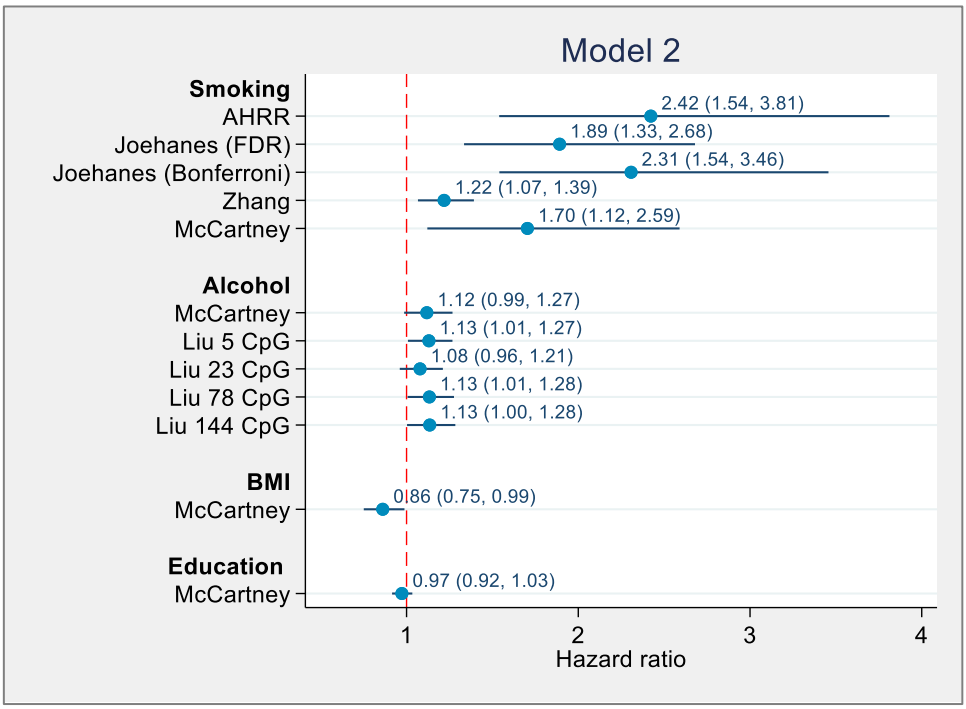
between alcohol-related DNAm scores with survival. The estimated effects of the alcohol, BMI, and education DNAm scores on mortality risk were broadly comparable to those obtained in the minimally adjusted model.

Figure 44: Forest plots showing the estimated hazard ratios and corresponding 95% confidence intervals for the associations of DNAm predictors with all-cause mortality (n=408).

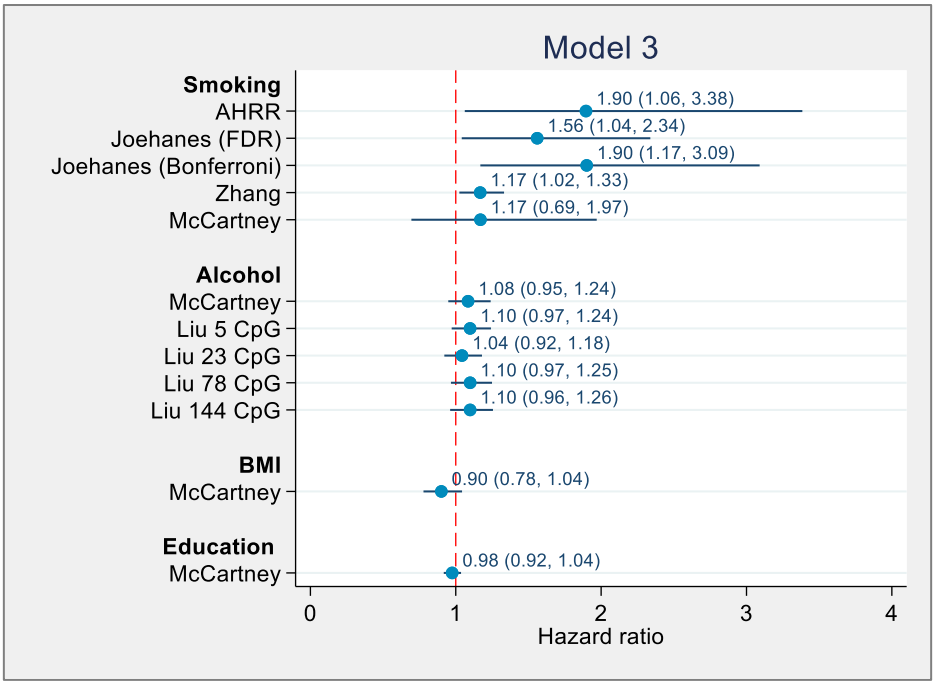
a. Adjusted for gender, age, cell counts and batch effects.



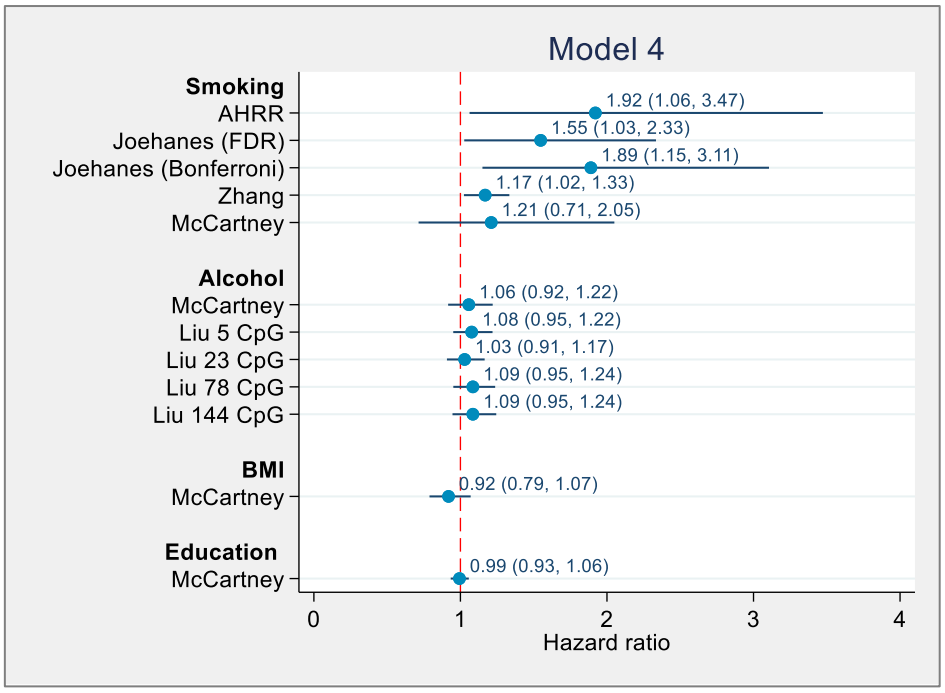
b. Additionally, adjusted for TNM stage, HPV status and comorbidity.



a. Additionally, adjusted for the respective directly measured phenotype.



b. Additionally, adjusted for the other directly measured phenotypes.



The different DNAm scores predicting smoking, alcohol drinking, BMI and education are displayed on the Y-axis. The X-axis shows the hazard ratio for all-cause mortality risk per SD unit increase in DNAm score. The 95% confidence intervals (CIs) are provided in the brackets. Abbreviations: ***BMI****, body mass index.;* ***CpG****, cytosine-guanine dinucleotides*.

8.5. Discussion

8.5.1. Principle findings

Genome-wide DNAm profiling has enabled the development of several biological predictors of complex traits including smoking, alcohol consumption, BMI and educational attainment. This chapter provides evidence that, in this cohort of individuals with OPC, epigenetic predictors of smoking not only explain a high proportion of the phenotypic variance - as much as 49% - they are also associated with survival following diagnosis. This was the case even after controlling for self-reported smoking, suggesting the DNAm predictors provide additional information on smoking history. The DNAm score developed by Joehanes *et al,* which applied a stringent Bonferroni threshold to select CpGs for inclusion, was consistently associated with all-cause mortality in both primary and sensitivity analysis. Compared to people with a mean (Joehanes) DNAm score for non-smokers, people with a mean (Joehanes) DNAm score for current smokers were around twice as likely to die during follow-up (fully adjusted HR= 1.89 [95% CI: 1.15, 3.11; p=1.2x10^{-2}] in the primary analysis and 2.11 [95% CI: 1.08, 4.12; p=2.99 x 10^{-2}] in the complete case analysis). The AHRR methylation score was also associated with around a two-fold increased all-cause mortality risk in the primary analysis (fully adjusted HR= 1.92 [95% CI: 1.06, 3.47; p=3.08 x 10^{-2}]). These results support the findings of my previous chapter, which found that self-reported current smokers had around a two-fold increased risk of death compared with self-reported non-smokers (for oral, oropharyngeal and laryngeal cancers combined), suggesting that DNAm based predictors for smoking could provide a useful tool in epidemiological studies to identify the smoking profiles of individuals in the absence of reliable self-report.

It was notable that the risk associated with smoking, as predicted by methylation status at AHRR, was as high as the risk associated by the Joehanes score; in fact, the effect estimate was slightly higher. This is important because it suggests that, in prognostic studies, a smoking proxy based on differential methylation at a single CpG site (described by Illumina array probe ID cg05575921) may be just as good, if not better, than a multi-probe score based on thousands of CpG sites across the genome. Prior work suggests that reversion of AHRR hypomethylation can provide a quantitative biomarker of smoking cessation [638], therefore future studies could use this epigenetic predictor to examine the prognostic impact of smoking cessation in HNC.

The DNAm predictors for alcohol consumption were associated with all-cause mortality in minimally adjusted models, and following adjustment for clinical factors, but associations were not robust to adjustment for self-reported drinking. Similarly, there was evidence that a higher BMI was protective in models 1 and 2, with HRs of 0.88 (95% CI: 0.77, 1.01; p=6.25 x 10^{-2}) and 0.86 (95% CI: 0.75, 0.99; p=3.27 x 10^{-2}) respectively, but on adjustment for self-reported BMI (model 3), there was little evidence supporting this association. There was little evidence to suggest that the DNAm based predictor for educational attainment was related to survival.

The DNAm predictors for alcohol consumption, BMI and education explained a smaller proportion of the variance in their phenotypic trait than the smoking DNAm scores in this analysis (7-16% for alcohol, 22% for BMI and 0.68% for education, respectively). These findings are line with those of an earlier analysis by McCartney *et al*, who used data derived from whole blood samples in a large cohort - Generation Scotland (a healthy population), to generate DNAm scores. The exception being that in their dataset, the proportion of phenotypic variance explained by their smoking predictor was almost 61% [633]. The power of the DNAm-based predictor for smoking was well demonstrated in their study, with near-perfect discrimination between current and never smokers based on ROC analysis (AUC=0.98). Together, these findings suggest that DNAm-based predictors of smoking could augment phenotypic prediction of mortality in future epidemiological studies. The same authors reported that models including polygenic scores (i.e. genetic scores) in addition to DNAm scores explain a greater proportion of variance in these traits [633], albeit it is still less than is explained for smoking. It would be interesting to see whether these combined predictors are associated with mortality risk in this cohort.

8.5.2. Strengths and limitations of the study

This study represents a novel application of epigenetic prediction of phenotypes within a cancer cohort. Two key strengths of the study include the availability of pre-established, robust epigenetic risk scores (DNAm scores) from large genome-wide consortia and the availability of MethylationEPIC data in H&N5000. The MethylationEPIC platform, which has been used in multiple published EWAS, including those used to derive the epigenetic risk scores employed in the current analysis, includes ~400,000 more CpG sites than the preceding 450k array (whilst maintaining the vast majority of sites already included on the 450K array). As such, the epigenetic risk scores used in this analysis explain a greater proportion of phenotypic variance than those derived using the earlier 450k array.

The availability of both epigenetic data and comprehensive all-cause mortality follow-up data in the same clinical cohort represents another major strength of this study, as does the ability to adjust for multiple biological, clinical and lifestyle covariates, including HPV, which are especially relevant to this disease.

As was the case with the observational analysis described in Chapter 6, a major limitation of this study is that it was not possible to estimate the association of exposures (here DNAm scores) with OPC-specific mortality, that is, I only had information on all-cause mortality. This affects how the results may be used clinically. For example, it is unclear whether the smoking association relates to cardiovascular outcomes only or if it also affects the progression of OPC. Another potential limitation of this study is that the number of participants with epigenetic data available was relatively small, meaning that the number of outcome mortality events was also small. This created multiple issues; firstly, there was limited statistical power to test the PH assumption and secondly there was limited power to detect the true effect size of any potential associations of epigenetic predictors with mortality, as reflected by the relatively large CIs for some of the smoking predictors. When the data were examined, it was noted that there were relatively few events after 3-years. In an attempt to address the PH issue, a sensitivity analysis was conducted in which the data were censored at 3-years. The results of the primary analysis and the censored analysis were similar in that only the smoking predictors were associated with mortality following full adjustment, and the HRs overlapped.

It is important to note that the aim of this analysis was to investigate whether the DNAm risk scores are associated with all-cause mortality; as such, these results do not establish whether DNA methylation explains the association of smoking with increased mortality. Further studies with larger sample sizes are needed to provide mechanistic insight.

8.5.3. Conclusions

In summary, the results of the present analysis suggest that DNAm based biomarkers of smoking exposure have the potential to improve exposure assessment in epidemiological and clinical settings when reporting bias is an issue and could be used to inform risk prediction.

Chapter 9: Associations of epigenetic biomarkers of ageing with mortality risk in oropharyngeal cancer

9.1. Introduction

We saw earlier in this thesis that the mortality rates for OPC vary considerably. The overall 5-year survival rate is around 50% [639] but estimates range from 35-83% [180 640]. As such, the ability to accurately predict an individual's survival probability at the time of diagnosis is important for clinical decision making and the enrolment of low-risk individuals into treatment de-escalation trials, which aim to achieve similar efficacy as routine clinical therapies but with reduced toxicity and improved quality of life [641].

HPV status has already been highlighted as an important prognostic factor for OPC [348] ; it is now included in prognostic models alongside TNM stage and comorbidity [418 483]. One such model yielded a Harrell's concordance statistic (C-Statistic) of 0.69 in external validation, indicating good (but not excellent) prediction [483]. The potential for model improvement is currently being explored and the prognostic value of various lifestyle factors such as smoking and alcohol intake have been investigated [441 442 446 450 575]. The prognostic role of epigenetic markers, including DNAm, are less well studied.

In Chapter 8, I started to explore whether DNAm could enhance outcome prediction in people with OPC. Specifically, I analysed the associations of DNAm-based predictors for smoking, alcohol intake, BMI and educational attainment - phenotypes that have all been linked to OPC outcomes in the literature, with survival. I found strong evidence that smoking-related DNAm scores may provide informative molecular biomarkers for mortality risk.

As well as being modified by lifestyle behaviours such as smoking, DNAm levels also vary considerably with age [527]. Models of epigenetic ageing ("epigenetic clocks"), which were first introduced in Chapter 4, have been implicated as potentially useful biomarkers for age-related disease and mortality, including risk of cancer [518 522]. To recap, epigenetic clocks comprise a set of CpG sites whose DNAm levels are known to increase (or decrease depending on the CpG site) with age, and which together predict chronological age with a high degree of accuracy. The majority of studies evaluating the ability of these epigenetic clocks to predict morbidity and mortality to date have been conducted in general (healthy)

populations [521] [527]; there are very few studies investigating their prognostic value in a clinical setting. One study used a Cox model to estimate HRs for the association between epigenetic age acceleration (EAA), that is the discrepancy between DNAm age as measured by the epigenetic clock and chronological age, and risk of death following cancer diagnosis (n=1,726 deaths) [642]. After adjusting for sociodemographic and lifestyle variables (smoking, alcohol intake, healthy eating, physical activity, socioeconomic status and education), the authors found limited evidence of an association with epigenetic age acceleration based on an epigenetic clock derived from methylation at 353 CpG sites (*EAAHorvath*) [523] but mortality risk was 10-30% higher for the highest versus lowest quartile of age acceleration based on an epigenetic clock derived from methylation at 71 CpG sites (*EAAHannum*) [520]. There was no evidence of heterogeneity by cancer type (breast, colorectal, gastric, kidney, lung, b-cell lymphoma, prostate, urothelial cell carcinoma).

Using several previously published methods for modelling epigenetic age, this chapter investigates the potential relationships between all-cause mortality and EAA in OPC. In particular, it assesses associations between both the “first generation” epigenetic clocks [520] [523], derived from DNA methylation levels at CpG sites found to be strongly associated with chronological age (i.e. *EAAHorvath* and *EAAHannum*), as well as more-recently derived clocks: one optimised to predict physiological dysregulation (*AgeAccelPheno*) [524] and one optimised to predict lifespan (*AgeAccelGrim*) [525].

9.2. Aims and objectives

The overall aim of this analysis was to assess whether EAA is associated with all-cause mortality in a sub-set of individuals with OPC in H&N5000. The objectives were:

1. To establish whether five existing measures of EAA (*EEAA, IEAA, IEAAHannum*, *AgeAccelPheno* and *AgeAccelGrim*) are associated with all-cause mortality, after controlling for established risk factors;
2. To explore whether the inclusion of these biomarkers of epigenetic aging improves mortality prediction compared to a clinical model based on age, tumour stage, HPV status and comorbidity.

9.3. Methods

9.3.1. Study population

The study population for this analysis were the 408 individuals with OPC, for whom epigenetic data, baseline questionnaire data and data capture were available (Figure 45). Epigenetic data was generated as per the previous chapter using the Infinium MethylationEpic Bead Chip (EPIC array). Full details are provided in the Methods Chapter (Chapter 4: *The Head and Neck 5000 study*). The methylation level at each CpG site was calculated as a beta value (β), which is the ratio of the methylated probe intensity and the overall intensity and ranges from 0 (no cytosine methylation) to 1 (complete cytosine methylation).

9.3.2. Estimation of epigenetic age

Epigenetic biomarkers of aging were first described in Chapter 3 (*Capturing exposures*). The epigenetic aging measures for this analysis were kindly provided by Dr Rebecca Richmond at the Integrated Epidemiology Unit (IEU), University of Bristol. To generate the epigenetic aging measures in H&N5000, DNA methylation data for a subset of 27,523 CpG sites from the Illumina EPIC array were uploaded on to the online DNA Methylation Age Calculator (https://dnamage.genetics.ucla.edu/) developed by Horvath's group, along with an annotation file containing data on chronological age, sex and tissue type for the samples. This subset of sites was chosen based on the list of 30,085 CpGs listed in the "datMiniAnnotation3.csv" file available for "Advanced Analysis" analysis on the DNA Methylation Age Calculator website. 2,562 of the CpG sites were missing due to probe discrepancy between the Illumina EPIC platform and Illumina 450K platform, the latter of which was used to derive some of the epigenetic clocks.

For each of the 408 individuals, the following epigenetic aging measures were generated (names follow the notation of previous publications):

- intrinsic epigenetic age acceleration based on Horvath's multi-tissue predictor (*IEAA*) [523];
- intrinsic epigenetic age acceleration based on Hannum's predictor (*IEAAHannum*) [520];
- extrinsic epigenetic age acceleration (*EEAA*), an enhanced version of Hannum's method, which up-weights the contribution of blood cell composition [522];
- PhenoAge (*AgeAccelPheno*) [524];
- and GrimAge (*AgeAccelGrim*) [525].

An overview of the different epigenetic age predictors is provided in Table 38. Intrinsic epigenetic age acceleration (*IEAA*) is defined as the residual resulting from a linear regression of estimated DNAm age, as predicted by the epigenetic clock, on chronological age and estimates of plasmablasts, naïve and exhausted CD8+ T cells, CD4+ T cells, natural killer cells, monocytes, and granulocytes (estimated from the methylation data). Extrinsic epigenetic age (*EEAA*), by comparison, is defined as *IEAA* plus a weighted average of three cell types that are known to change with age (naïve [CD45RA+CCR7+] cytotoxic T cells, exhausted [CD28-CD45RA-] cytotoxic T cells, and plasmablasts) and by this definition is able to capture aspects of immunosenescence [522].

Table 38: Overview of various measures of epigenetic age acceleration.

Measure	Abbreviation	CpGs	Description
Intrinsic epigenetic age acceleration based on Horvath	*IEAA*	353	The residual resulting from regressing DNAm age on chronological age and estimates of major blood immune counts.
Intrinsic epigenetic age acceleration based on Hannum	*IEA Hannum*	71	
Extrinsic epigenetic age accel. based on Hannum	*EEAA*	71	Residual resulting from a univariate model regressing a weighted age estimate (which increases the contribution of 3 cell types known to change with age) on chronological age.
Age acceleration based on PhenoAge	*AgeAccelPheno*	513	Residual resulting from a linear model regressing AgePheno on chronological age, where PhenoAge is a measure based on a linear combination of chronological age and 9 clinical biomarkers (albumin, creatinine, glucose, serum, CRP, lymphocyte percent, mean cell volume, red cell distribution width, alkaline phosphatase, white blood cell count, age).
Age acceleration based on GrimAge	*AgeAccelGrim*	1,030	Residual resulting from a linear model regressing GrimAge on chronological age, where GrimAge is a measure based on a linear combination of chronological age, sex, and DNAm-based surrogate biomarkers for smoking pack-years and seven plasma protein levels (ADM, B2M, cystatin C, GDF-15, leptin, PAI-1,TIMP-1).

*Abbreviations: **ADM**, adrenomedullin levels; **B2M**, beta-2 microglobulin; **CRP**, C-reactive protein; **EAA**, epigenetic age acceleration; **GDF-15**, growth differentiation factor 15; **PAI-1**, plasminogen activation inhibitor 1; **TIMP-1**, tissue inhibitor metalloproteinase 1.*

9.3.3. Statistical analysis

Stata (Release 15.1, StataCorp) was used for all analyses described below. The study was split into two steps (Figure 45). The first step explored the potential associations of EAA measures with survival, after controlling for established HNC prognostic factors (listed below); The second step investigated whether these EAA measures provide any additional prognostic information, over and above those factors that are routinely considered (and which are available) in clinical practice, namely age, gender, tumour stage, HPV status and comorbidity .

9.3.3.1. Step 1: Examining the association of EAA measures with survival

Descriptive analyses were performed to explore the distribution of, and correlations between EAA measures, using histograms and Pearson's correlation coefficients. Baseline descriptive data were stratified by whether or not they were alive at 3-years. The univariate association of covariates on all-cause mortality risk was assessed using Kaplan-Meier curves and the log-rank test.

Multivariable Cox proportional hazards models were used to examine the potential associations of the epigenetic age measures with all-cause mortality. Again, cancer-specific mortality data were not available at the time of data analysis. Given that the five epigenetic measures of age acceleration are expressed in different units, measures were standardised using z-scores to allow comparison of effect estimates. To calculate the z-score, I computed the difference between a given value and the mean for that measure and divided it by the standard deviation (SD). So, a z-score of 0 would be equal to the mean, a z-score of 1 is 1 SD above the mean and a z-score of -1 is 1 SD below the mean. HRs and 95% CIs for all-cause mortality were calculated for each SD increase in EAA. For each measure of epigenetic aging, four separate Cox models were run:

1) a minimally adjusted model that only controlled for gender;
2) a model that additionally controlled for clinical factors (TNM stage, HPV status, comorbidity and BMI);
3) a model that additionally controlled for socioeconomic factors (education, annual house-hold income, marital status);
4) a fully adjusted model that additionally controlled for modifiable lifestyle behaviours (smoking and alcohol drinking).

Models were selected *a priori* based on the existing literature linking these covariates with survival [388 441 643-646] (See chapter 2).

Several of the covariates of interest had some missing data (details provided in Chapter 7; Table 30), particularly BMI as this measure was not initially collected at the start of recruitment into the H&N5000 study. Excluding individuals with missing covariate data would have reduced the statistical power to detect an association between our exposures of interest and survival, and so MI was performed. As mentioned previously, earlier work suggests that MI provides unbiased results in situations where data are missing at random [647] (see Chapter 7). Missing values were imputed using the ICE package for multiple chained equations in Stata [613]. Twenty imputed datasets were generated and analysed separately using standard statistical methods and the multiple sets of results combined using Rubin's rules [614], as per the previous chapter. The imputation models contained all the variables in the analysis model (including the outcome) and Nelson–Aalen estimator of the cumulative hazard. No outcomes were imputed. As a sensitivity analysis, a complete case dataset, which only included those participants with data available for all of the covariates of interest, was also generated and analysed as above [615].

9.3.3.2. Step 2: Assessing the prognostic value of EAA measures

Evidence of an association with outcome is not enough to include novel biomarkers in prediction models; to be useful to clinicians they must provide added prognostic value to existing models [648]. Therefore, step 2 of the analysis explored whether the addition of EAA measures to models based on established mortality risk factors (i.e. those currently considered in clinical decision making), improve model performance.

Flexible parametric survival models were fitted using the methods of Royston and Parmar [649 650], which model the baseline hazard (on the log cumulative hazards scale) using restricted cubic splines [651]. Splines are mathematical functions formed by piecewise polynomials [652]. They are fit with some constraints to ensure that the overall curve of the baseline distribution is smooth, meaning that more complex shapes can be fit. Flexible parametric models are conceptually very similar to the Cox model, and they provide similar estimates, but they have certain benefits if you are interested in prediction. Unlike Cox regression, flexible parametric models permit absolute (as opposed to relative) measures of effect (i.e. survival probability) to be estimated at all time points, rather than just at event times [653]. To put this another way, using the flexible parametric model structure, outcomes can be predicted for time points

other than those that the prognostic model was developed on (see Riley, 2019 for calculations [654]). Time dependent effects (non-proportional hazards) can also be modelled using this approach, which is particularly useful when you are working with a wide range of variables (demographic, clinical and biological). This is because, as the number of variables in your model increases, the probability that one or more of these variables fails to satisfy the PH assumption becomes increasing large. Indeed, there was a suggestion in the previous chapter that the PH assumption might not be reasonable for HPV status (see Chapter 7; Supplementary figure 2).

The Royston and Parmar (RP) models were fitted using maximum likelihood estimation via the `stpm2` command in Stata. The spline complexity for the baseline hazard function which best fits the data was investigated visually and through model fit statistics (Akaike Information Criterion [AIC] and Bayesian Information Criterion [BIC]). The AIC is defined as 2Log(likelihood) + 2(No. of model parameters), while BIC equals 2Log(likelihood) + (No. of model parameters) *Log(n). Degrees of freedom (df) ranging from 1 to 5 df (for a model with no variables included) were considered. Using the hazard function plots and the AIC and BIC as a guide, 2 df were deemed sufficient. These 2 degrees of freedom equate to 1 interior knot, the point where the polynomials join [655]. Non-linear relationships with continuous predictors were considered using the multivariable fractional polynomial (MFP) algorithm described by Sauerbrei and Royston [656] and implemented in Stata using the `mfp` command, which selects the MFP model that best predicts the outcome variable.

The following models were fit:

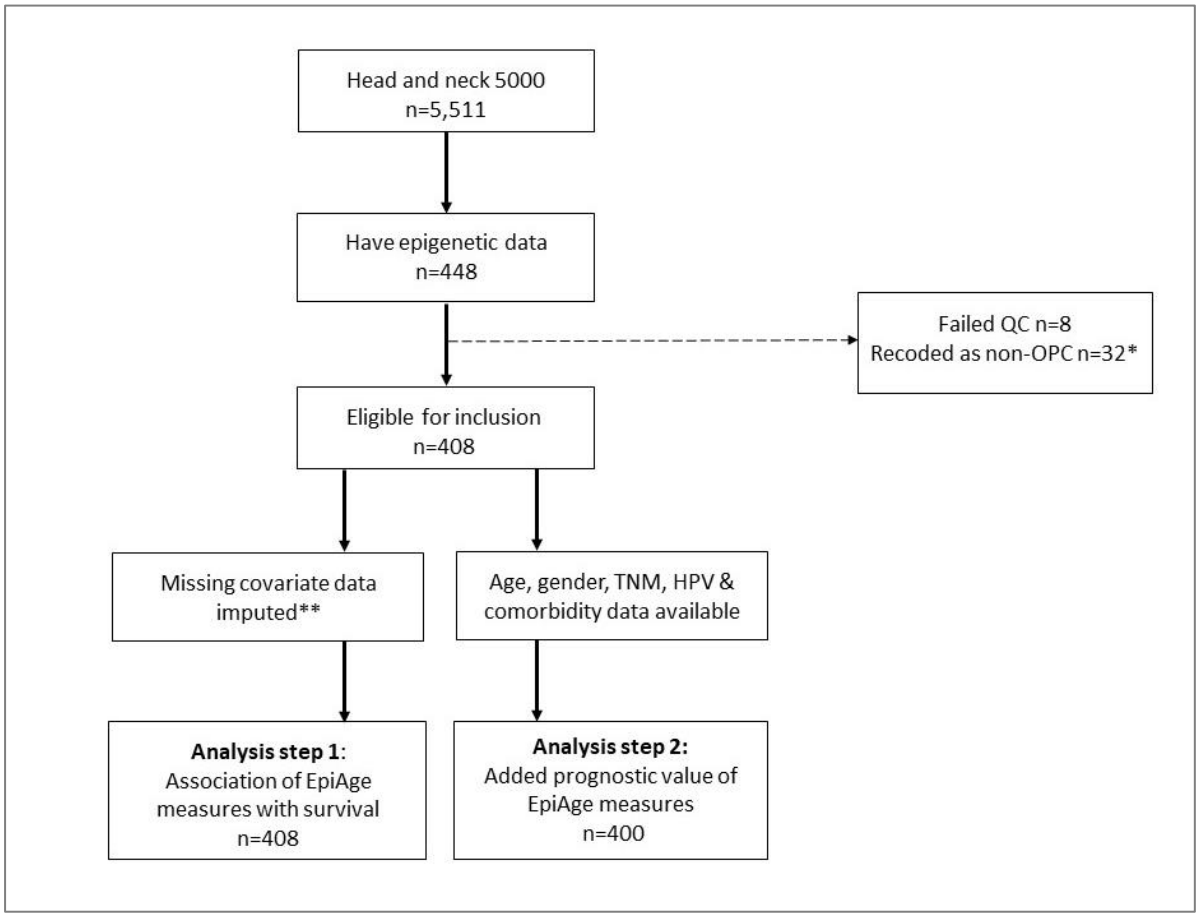
1) a ‘clinical model’, which comprised age, gender, TNM stage, HPV status and comorbidity;
2) clinical + *IEAA*;
3) clinical + *EEAA*;
4) clinical + *IEAAHannum*;
5) clinical + *AgeAccelGrim;* and
6) clinical + *AgeAccelPheno.*

Models were fit in a sub-sample of participants with complete data available for the clinical covariates considered in this analysis (age, gender, tumour stage, comorbidity, and HPV status). The performance measures examined were the AIC, which measures the relative goodness of fit of a model, considering both the statistical goodness of fit and the number of parameters used, and the C-statistic, an extension of the area under the receiver operating curve (AUC) to survival analysis [465 657]. The interpretation of the statistic is equivalent,

namely, a C-statistic of 0.5 indicates no discrimination above chance (of dying or surviving), whereas a C-statistic of 1.0 indicates perfect discrimination, and thus superior prediction [658].

ROC curves and AUC functions were calculated to characterize how well the fitted models distinguish between participants who were and were not alive three years after diagnosis. Internal validation was performed on the final model using 500 bootstrap samples to adjust performance for optimism and calculate a shrinkage factor to be applied to the models' regression coefficients for use in other (external) settings [659] [660] .

Figure 45: Flow of participants included in the Epigenetic Age analysis.



**Samples were initially selected for genetic and epigenetic analysis based on clinical diagnosis of OPC (ICD-10: CO1, CO5, CO9, C10.0-2, C10.3, C10.8 and C10.9), as recorded by the study Centre. Pathology reports of individual cases were subsequently checked (where possible) to verify tumour site. **Data available for age, gender, TNM stage, HPV status, comorbidity, education, BMI, self-reported smoking status and alcohol consumption.*

9.4. Results

A total of 105 deaths were observed during follow-up (median=5.3 years, IQR: 4.9 to 6.0; n=408). The proportion of missing data in this data set was summarised in the previous Chapter (Table 30). Information on age, stage and HPV status were complete. The largest proportion of missing data corresponded to BMI (33.3%), followed by annual household income (12.5%) and education (4.7%). The complete case sample included 225 participants (n=49 deaths).

9.4.1. Baseline descriptives

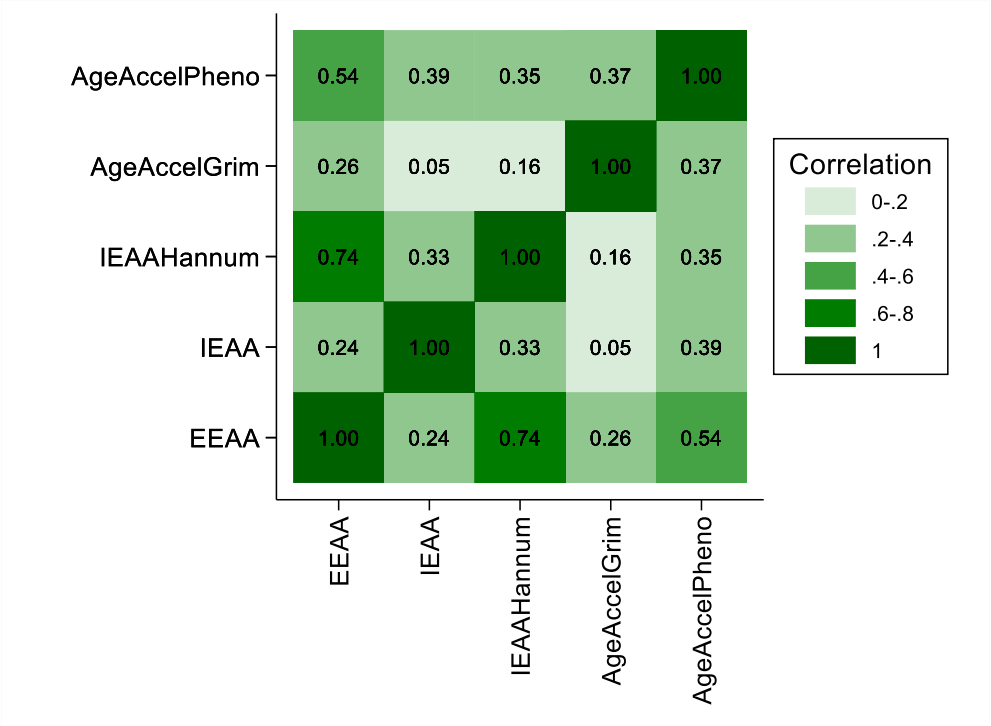
The baseline clinical and sociodemographic characteristics of participants included in the epigenetic dataset stratified by tumour site and HPV status were provided in Chapter 6 (Tables 21 and 22). Table 39 presents the baseline characteristics and mean EAA measures of participants stratified by whether or not they were alive at 3-years. There were differences in TNM stage, HPV status, comorbidity, annual household income and smoking and alcohol intake across groups. Compared to the group who were alive at 3-years, the group who had died were more likely to have been diagnosed with stage III or IV cancer (94% versus 85%), were more likely to be HPV-negative (58% versus 23%), were more likely to have moderate/severe comorbidity (30% versus16%), were more likely to be earning <£18,000 (56% vs. 35%), were more likely to be current smokers (34% versus 16%) and were more likely to drink hazardous to harmful amounts of alcohol (66% versus 48%). The average age of participants was 63 years in the group who had died and 57 years in the group that were alive. Mean EAA, as measured by each of the epigenetic clocks, was higher in those individuals who had died at three years. There were differences between people who had survived and those who had died for the following epigenetic age measures*: EEAA* (p=0.004), *IEAAHannum* (p=0.006), *AgeAccelPheno* (p=<0.001) and *AgeAccelGrim* (p=0.002). Only IEAA showed no association with survival status at 3 years.

The distribution of baseline characteristics for participants included in the complete case analysis was broadly comparable with respect to age, gender, HPV status, smoking status, alcohol intake, education, income and marital status (Appendix C1). However, there was no evidence of differences in comorbidity (p=0.204) or TNM stage (p=0.740) between people who were and were not alive at 3-years in the complete case analysis.

9.4.2. Pairwise correlations between measures of epigenetic age acceleration

The pairwise Pearson correlation coefficients of selected measures of *EEAA* is shown in Figure 46. The measures exhibited correlation coefficients ranging from r=0.05 to r=0.74. The strongest absolute association was between *EEAA* and *IEAAHannum*.

Figure 46: Heat map showing the pairwise correlation coefficients between epigenetic measures of age acceleration.



The heatmap colour-codes the pairwise Pearson correlation coefficients between epigenetic measures of age acceleration based on the multi-tissue epigenetic clock developed by Horvath (IEAA) [523], the blood-based DNAm age developed by Hannum (IEAAHannum and EEAA) [520 522], DNAm GrimAge (AgeAccelGrim) developed by Lu [525], and DNAm PhenoAge developed by Levine (AgeAccelPheno) [661]

Table 39: Baseline characteristics of the study sample stratified by 3-year mortality status (n=408).

	Dead at 3 years (n=77)		Alive at 3 years (n=331)		
Characteristic	N	%	N	%	p-value
Gender					
Male	60	77.90%	257	77.60%	
Female	17	22.10%	74	22.40%	0.958
TNM stage					
I	1	1.30%	16	4.80%	
II	4	5.20%	35	10.60%	
III	14	18.20%	44	13.30%	
IV	58	75.30%	236	71.30%	0.175
TNM stage group					
Low	5	6.50%	51	15.40%	
High	72	93.50%	280	84.60%	0.041
HPV status					
Negative	45	58.40%	77	23.30%	
Positive	32	41.60%	254	76.70%	<0.001
Comorbidity status					
None	26	34.20%	185	56.20%	
Mild	27	35.50%	92	28.00%	
Moderate/severe	23	30.30%	52	15.80%	0.001
Smoking					
Never	8	11.00%	102	32.00%	
Former	40	54.80%	165	51.70%	
Current	25	34.20%	52	16.30%	<0.001
Alcohol					
Non-drinker	14	18.90%	90	27.60%	
Moderate	11	14.90%	79	24.20%	
Hazardous/harmful	49	66.20%	157	48.20%	0.019
Education					
School education	37	50.00%	133	42.20%	
College	28	37.80%	130	41.30%	
Degree	9	12.20%	52	16.50%	0.422
Annual household income					
<£18,000	36	56.30%	102	34.80%	
£18000-£34,999	13	20.30%	90	30.70%	
>£35,000	15	23.40%	101	34.50%	0.006
Marital status					
single (never married)	11	14.70%	36	11.00%	
currently in relationship	38	50.70%	242	74.00%	
No longer with spouse	26	34.70%	49	15.00%	<0.001
	N	mean (SD)	N	mean (SD)	p-value
Age at baseline	77	62.86 (11.25)	326	57.39 (8.91)	<0.001
Body mass index	46	24.33 (4.76)	226	26.88 (4.87)	0.001
EEAA	77	1.68 (6.52)	331	-0.42 (5.53)	0.004
IEAA	77	0.36 (4.34)	331	-0.17 (4.38)	0.333
IEAAHannum	77	1.10 (4.52)	331	-0.27 (3.76)	0.006
AgeAccelerationResidualHannum	77	1.44 (5.04)	331	-0.35 (4.17)	0.001
AgeAccelPheno	77	3.16 (5.37)	331	-0.86 (5.40)	0.000
AgeAccelGrim	77	2.00 (7.04)	331	-0.62 (6.36)	0.002

Abbreviations: ***EEAA,*** *extrinsic epigenetic age acceleration;* ***IEAA****, intrinsic epigenetic age acceleration. P-value for difference based on the Chi-Square test (categorical) and one-way ANOVA (continuous). *Comorbidity was defined using the Adult comorbidity Evaluation Index-27 (ACE-27). All DNAm scores represent raw (i.e. non-z-scored) values.*

9.4.3. Explained variation in survival

The R_2 statistics for explained variation in survival are presented in Table 40. As with the previous chapter, which used the same dataset, the variance in survival explained by the basic model (age, gender, cell counts and batch effects) was 24%. The only EAA measure to enhance the proportion of explained variation was *AgeAccelGrim.* The addition of *AgeAccelGrim* to the model led to a 12% increase in explained variation (R_2=0.36 [95% CI:0.25, 0.45]).

Table 40: Proportion of variance in survival explained by the EAA measures (n=408).

Model	R^2	Lower	Upper
Basic model*	0.24	0.14	0.35
Additionally adjusted for DNAm predictors:			
IEAA	0.20	0.10	0.31
IEAA Hannum	0.24	0.14	0.34
EEAA	0.21	0.11	0.31
AgeAccelpheno	0.21	0.11	0.32
AgeAccelGrim	0.36	0.25	0.45

*Includes age, gender, cell counts (CD4 T-cells, CD8 -cells, eosinophils, monocytes neutrophils natural killer) and batch effects.

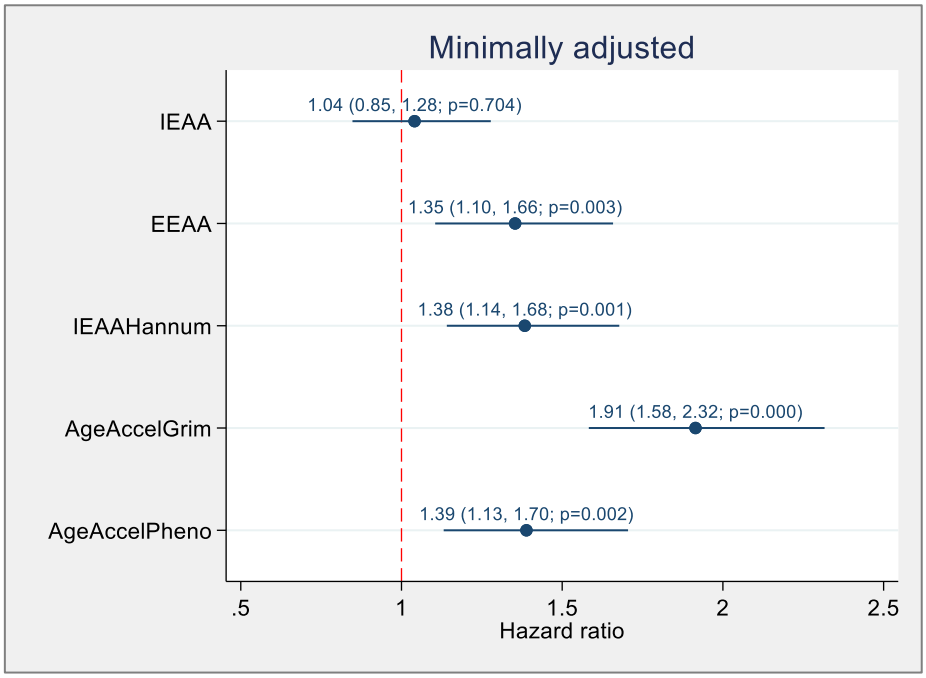
9.4.4. Association of DNA Methylation-Based Biological Age with survival

The results of the Cox regression analysis (imputed data; n=408) are presented in Figure 47. In the basic model (Figure 47 a.), which adjusted only for sex all of the epigenetic biomarkers of aging except the intrinsic epigenetic age acceleration measure *IEAA* were associated with survival. All of the reported associations are in the expected directions, i.e. higher values of EAA are associated with higher mortality risk. HRs ranged from 1.35 (95% CI: 1.10, 1.66; p=3.5 x10^{-03}) for *EEAA* to 1.91 (95% CI: 1.58, 2.32; p=2.3 x10^{-11}) for *AgeAccelGrim*, where HRs represent the difference in mortality risk per SD unit increase in EAA measure. With the exception of *AgeAccelPheno,* associations remained but were attenuated slightly following adjustment for clinical and socioeconomic factors (Figures 47 b. and 47 c.). In the fully adjusted model (Figure 47 d.), which also adjusted for smoking and alcohol intake, *IEAAHannum* and *AgeAccelGrim* were associated with mortality risk (HRs: 1.32 (95% CI: 1.08, 1.61; p=6.9 x10^{-03}) and 1.39 (95% CI: 1.06, 1.83; p=0.017), respectively,

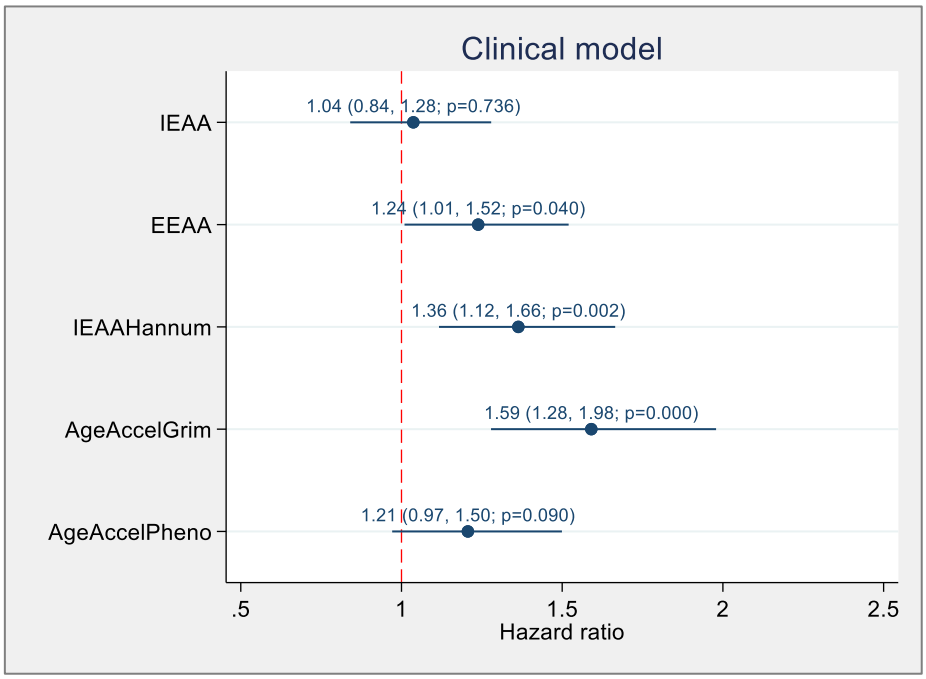
In the complete case analysis (n=225), the results of the basic model were broadly comparable to those of the imputed analysis (Appendix C2). However, *IEAAHannum* was not robust to adjustment for socioeconomic factors and the association of *AgeAccelGrim* with survival attenuated following adjustment for smoking and alcohol intake.

Figure 47: Association of epigenetic age acceleration measures with mortality risk.

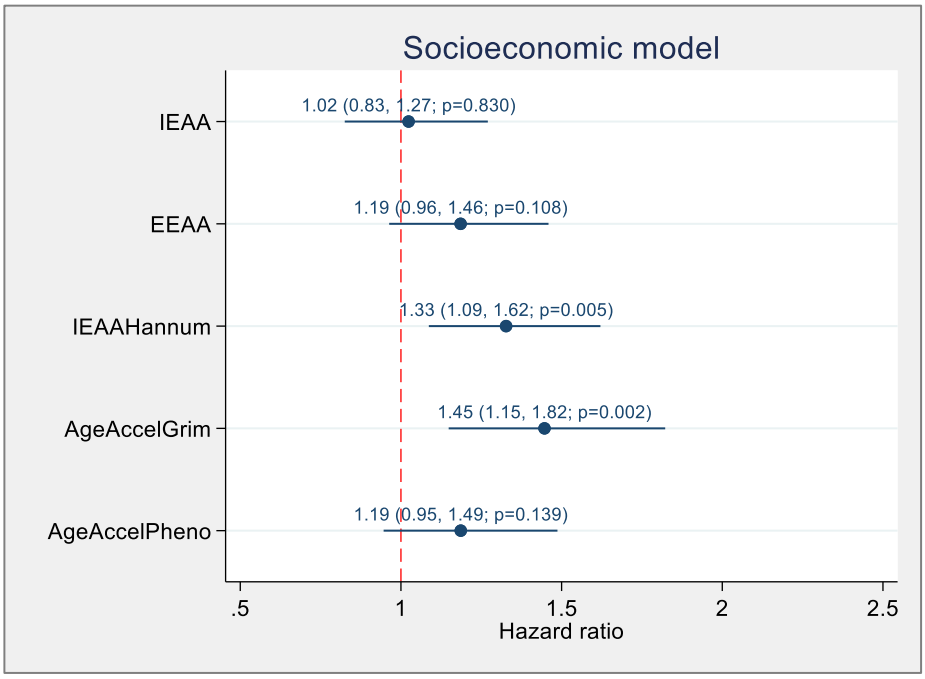
47 a: Adjusted for gender.



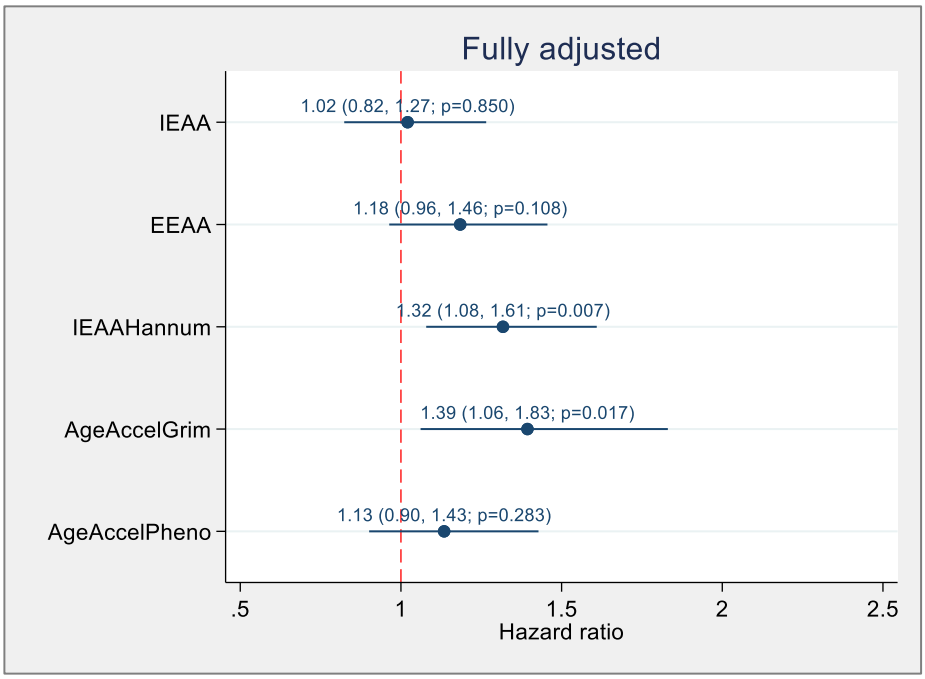
47 b: Additionally adjusted for tumour stage (high (III and IV) versus low (I and II) stage), HPV status, comorbidity and BMI.



47 c: Additionally adjusted for income, education and marital status.



47 d: Additionally adjusted for self-reported smoking status and alcohol intake.



Solid blue circles indicate the hazard ratios per SD increase of each of the age acceleration measures displayed on the y-axis. The corresponding 95% confidence intervals are shown in brackets. The red dashed line denotes a hazard ratio of 1, signifying lack of association with survival.

9.4.5. Examination of the predictive utility of EAA measures

Table 41 shows the performance measures for the fitted models. The AIC values for the clinical + *EEAA*, clinical + *IEAAHannum* and clinical + *AgeAccelGrim* model were lower than the clinical model. As a rule of thumb, two models are generally considered equivalent if the difference in their AICs is less than 2 units [662]; therefore on this basis, all three of these models had a better overall fit than the standard clinical model. The preferred model, i.e. the model with the lowest AIC value, was the clinical + *AgeAccelGrim* model. Compared to the clinical model (0.73), the C-statistics were higher for the clinical + *EEAA* (0.76), clinical + *IEAAHannum* (0.74) and clinical + *AgeAccelGrim* (0.76) models.

Given that the clinical + *AgeAccelGrim model* showed the strongest association in the Cox regression analysis, appeared to fit the data best and yielded the highest discrimination (i.e. had the highest AIC and C-statistic), I examined whether this model provided improved prediction at three years compared to a clinical model (including age, gender, TNM stage, HPV and comorbidity), by comparing AUC values. The results are illustrated in Figure 48. There was weak evidence to suggest the clinical + *AgeAccelGrim* model had superior predictive performance compared to the clinical model (clinical AUC: 0.77, clinical + *AgeAccelGrim* AUC: 0.80; *p*-value for difference=0.069), at which point there had been 76 deaths. The bootstrap optimism corrected AUC values showed a small reduction in performance compared with the original model (optimism adjusted AUCs of 0.74 and 0.77 for clinical and clinical + *AgeAccelGrim* models, respectively).

The optimism-adjusted c-slope (or uniform shrinkage factor) for the *clinical + AgeAccelGrim* model, was 0.83, indicating there was some overfitting. The original predictor effects adjusted by this value [658]. The results are presented in Table 42. In the adjusted model, each SD unit increase in *AgeAccelGrim* was associated with a 1.5-fold increased risk of death at 3 years (optimism adjusted HR= 1.54, 95% CI: 1.2, 1.92; p=<0.001).

Smoking has been shown to be independently predictive of mortality in H&N5000 [575]. The reduced effect estimate observed between *AgeAccelGrim* and mortality with adjustment for smoking status suggests that the enhanced prognostic ability gained from adding *AgeAccelGrim* to the clinical model could be due to the inclusion of a smoking predictor [525]. In order to investigate this, I conducted an additional sensitivity analysis whereby I compared the prognostic ability of the following models: 1) clinical + *AgeAccelGrim*; 2) clinical + self-

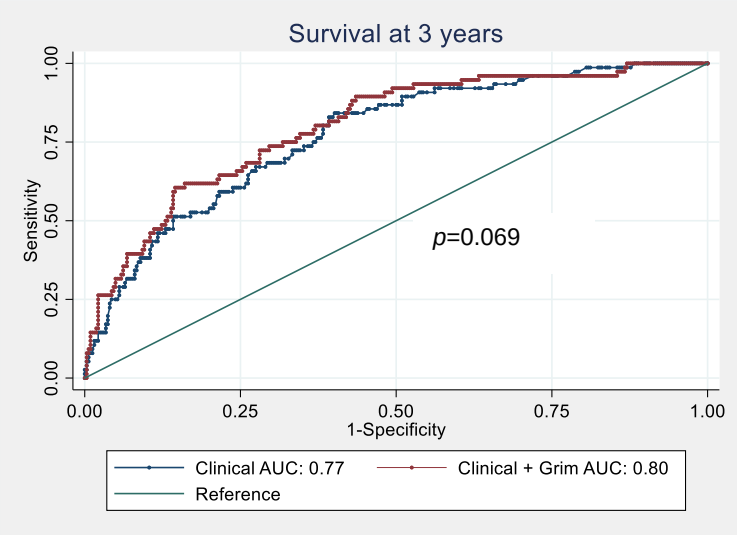
reported smoking; and 3) clinical + *dnampackyears*, the DNAm-based surrogate biomarker for pack years of smoking used to derive GrimAge (n=384 participants with smoking data available; no. deaths=72). At 3-years, there was a suggestion that the clinical + *AgeAccelGrim* model had better discrimination (AUC value of 0.80) than the clinical models including both self-reported smoking (AUC=0.77) and a DNAm surrogate for pack years (AUC=0.78) (Figure 49), although there was limited evidence of a difference in AUCs based on chi-squared tests (*p*=0.175).

Table 41: Measures of model performance.

Model	AIC	*C*-statistic (95% CI)
Clinical	486.93	0.75 (0.70, 0.80)
Clinical + *EEAA*	483.36	0.76 (0.71, 0.81)
Clinical + *IEAA*	488.14	0.76 (0.71, 0.81)
Clinical + *IEAAHannum*	480.10	0.77 (0.72, 0.82)
Clinical + *AgeAccelGrim*	473.14	0.78 (0.73, 0.83)
Clinical + *AgeAccelPheno*	485.52	0.76 (0.71, 0.81)

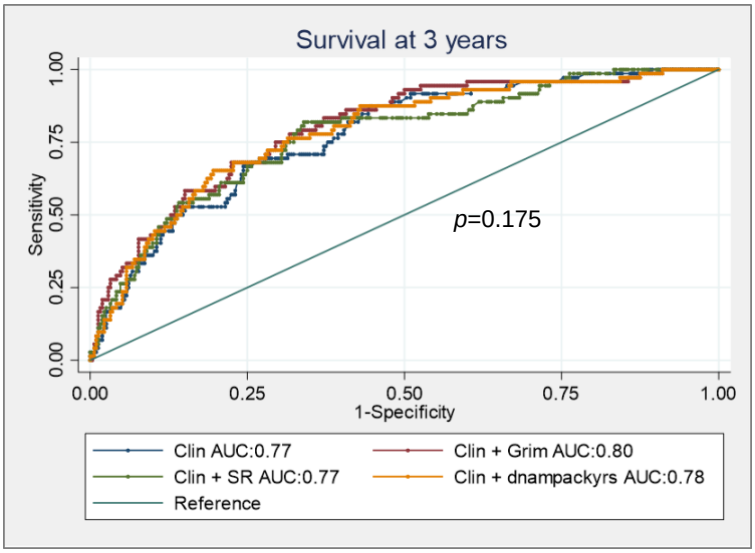
Abreviations: ***AgeAccelGrim****, age acceleration based on DNAmGrimAge;* ***AgeAccelPheno****; age acceleration based on PhenoAge;* ***AIC****, Akaike's information criterion;* ***C-statistic****, Harrel's concordance statistic;* ***EEAA****, extrinsic epigenetic age acceleration;* ***IEAA****, intrinsic epigenetic age acceleration;* ***95% CI****, 95% confidence interval.*

Figure 48: Independent contribution of AgeAccelGrim to prognosis beyond clinical factors (n=408).



Chi-squared p-values. Number of deaths at 3-years = 76.

Figure 49: A comparison of the area under the ROC curves (AUC) obtained for the models included in the sensitivity analyses (n=384).



Number of deaths at 3-years =72. Abbreviations: ***AUC****, area under the receiver operating characteristics curve;* ***Clin****, clinical model (including age, gender, TN|M stage, HPV status and comorbidity);* ***dnampackyrs****, the DNA methylation based biomarker of pack years of smoking used to derive GrimAge,* [525]*;* ***Grim****, age acceleration based on the GrimAge measure* [525]*;* ***SR****, self-reported smoking status*

Table 42: Estimated hazard ratios (uncorrected and corrected) for the clinical + AgeAccelGrim model.

	Original model			Final model after adjustment for overfitting		
		95% CI			95% CI	
Variable	HR	upper	lower	HR	upper	lower
Age	1.05	1.07	1.02	1.04	1.06	1.02
Gender						
Female	1.53	2.69	0.87	1.42	2.28	0.89
Tumour stage						
II	1.90	17.24	0.21	1.71	10.63	0.27
III	5.21	39.90	0.68	3.94	21.32	0.73
IV	6.36	46.75	0.90	4.65	24.31	0.89
HPV status						
Positive	0.39	0.65	0.23	0.45	0.70	0.29
Comorbidity*						
mild	1.39	2.46	0.80	1.32	2.11	0.82
moderate/severe	1.27	2.34	0.69	1.22	2.02	0.73
AgeAccelGrim	1.69	2.19	1.30	1.54	1.92	1.25

To obtain the adjusted estimates, the original model estimates (beta coefficients) were multiplied by the shrinkage factor (0.83) and exponentiated to convert to HRs. The beta coefficients are presented in Appendix C3.

9.5. Discussion

9.5.1. Principle findings

Age-related changes in DNAm have attracted growing attention in recent years, largely due to their ability to predict associated morbidity and mortality among general populations. The purpose of this chapter was to establish whether DNAm-based predictors of aging could provide useful biomarkers of mortality risk in people with OPC. The main findings were that two different epigenetic age estimators, *IEAAHannum* and *AgeAccelGrim,* were associated with increased risk of all-cause mortality and these associations were independent of

established mortality risk factors. *AgeAccelGrim* had the strongest effect estimate, with each SD increase in epigenetic age acceleration resulting in a 39% increase in risk of death in the fully adjusted model (HR=1.39; 95% CI: 1.06, 1.83). When compared to a standard clinical model that included age, sex, tumour stage, HPV status and comorbidity, the addition of *AgeAccelGrim* to the model showed some improvement in mortality risk prediction at 3-years (optimism-adjusted AUCs: 0.77 and 0.80 [*p*=0.069] for clinical and clinical +*AgeAccelGrim* models, respectively).

These findings support the current literature, which suggests that DNAm derived "GrimAge", a composite biomarker incorporating DNAm-based surrogates for smoking pack-years and seven plasma proteins in addition to age and sex, is a better predictor of mortality risk compared to the first-generation DNAm-based predictors (i.e. Horvath and Hannum's clocks) [525]. Age acceleration as measured by GrimAge has not only been shown to perform better in predicting time to death and time to cancer among general populations, it has also been associated with established cancer risk factors [525].

It is possible that *AgeAccelGrim* was most strongly related to mortality risk in the current analysis due to the inclusion of the surrogate measure for smoking in the GrimAge biomarker, since smoking has been shown to be independently predictive of mortality among HNC cases [575]. When the prognostic performance of the clinical + *AgeAccelGrim* model was compared with clinical models including both self-reported smoking and the DNAm surrogate biomarker for pack years of smoking, the clinical + *AgeAccelGrim* model had the best discrimination (AUC:0.80), followed by the clinical + dnampackyrs (AUC:0.78), though the improvement in predictive performance was marginal. These findings suggest that the methylation-based measure of smoking is a better indicator with less misclassification than self-report and that the prognostic utility of *AgeAccelGrim* does not appear to be driven solely by the inclusion of the DNAm-based biomarker for smoking.

As mentioned, GrimAge is also trained on a set of proteins known to be associated with mortality [525], including plasminogen activator inhibitor 1 (PAI-1) and growth differentiation factor 15 (GDF15). PAI-1 (aka SERPIN E1) is overexpressed in a variety of tumours and has been found to be a strong predictor of poor clinical outcome and poor response to therapy [663-665] whilst GDF15 is involved in the pathogenesis of oral squamous cell carcinoma (OSCC) [666-668]. Several studies have demonstrated that, in people with OSCC, GDF15 is also associated with decreased survival and diminished response to chemotherapy [666 667 669 670]. The inclusion of DNAm-based surrogate measures for these proteins may go part way to

explaining the enhanced prognostic ability of *AgeAccelGrim* in the current analysis and should be investigated further.

9.5.2. Strengths and limitations of the study

This investigation has several strengths including the relatively long follow-up period, the fact that individuals were sampled at the time of diagnosis and the fact that DNAm was assayed simultaneously in the same laboratory. I was also able to account for a range of factors that could confound the effect estimates, including smoking, alcohol intake and BMI, all of which are known to influence DNAm and HNC risk [633 671]. Moreover, missing covariate data were imputed via chained equations to minimise possible biases [610 672].

Several limitations should be highlighted. First, the sample size for this analysis was small and I was unable to identify an independent prospective dataset to validate my findings. In an attempt to mitigate possible optimism in my effect estimates, I calculated a uniform shrinkage factor and multiplied this by the original β coefficients from the fitted model. I then presented the optimism-adjusted coefficients. Second, information on some of the variables used in this analysis were obtained via participants' self-report, which earlier chapters explained can result in recall bias or misreporting. I employed a DNAm-derived measure of packyears of smoking in my sensitivity analysis but future studies could implement the use of other methylation scores to index these variables (e.g. BMI and alcohol intake) [633 671]. Fourth, there is a disparity in coverage between 450K and 850K Illumina platforms meaning that 19 of the 353 CpGs included in Horvath's clock and 6 of the 71 CpGs included in Hannum's clock are missing in H&N5000. This could be problematic, although a previous study examining the application of EPIC array data to predict DNAm age demonstrated that the lack of the clock-CpGs on the EPIC array did not undermine the utility of the epigenetic age predictors [673]. Finally, I did not account for multiple testing of the 5 epigenetic age acceleration measures, although evidence of correlation between some of the epigenetic measures suggests that correction may not have been appropriate.

9.5.3. Conclusions

Here, I investigated the relationship between epigenetic measures of age acceleration and overall survival in blood samples of people with OPC. Overall, my findings provide evidence that DNA methylation-based estimators of ageing could provide prognostic utility, above established prognostic factors including age, sex, tumour stage, HPV status, comorbidity,

and smoking. That an accurate and easy-to-measure biomarker derived from peripheral blood could serve as a better predictor of mortality risk in people diagnosed with OPC is important as this could impact treatment planning and provide information that improves patient stratification in study design, e.g. treatment de-escalation trials. Nonetheless, these findings should be further investigated in a larger, independent sample.

Chapter 10: Metabolic signatures of oropharyngeal cancer survival

10.1. Introduction

The last two chapters of my thesis focused on epigenetic biomarkers and their ability to predict survival in people with OPC. They provided evidence that DNAm-based estimators of smoking and biological aging are associated with mortality and could improve prediction compared to models based on conventional HNC prognostic factors, including age, tumour stage and HPV status. Attention now turns towards the metabolome and the identification of metabolic predictors of survival in this cohort.

The metabolome, as described in Chapter 3, represents a functional readout of both upstream omics profiles (genomics, transcriptomics, proteomics) and exogenous environmental exposures. It provides an indication of what is happening in a cell, tissue or biofluid at that moment in time. In this respect, the metabolome may be considered more proximal to the phenotype than the genome or the proteome. Through studying metabolite changes that occur in response to a particular disease state or pathological phenotype, it may be possible to gain insight into the pathogenesis and progression of that disease and potentially aid in the development of novel diagnostic and prognostic biomarkers.

The majority of the metabolic processes in the body, such as those responsible for the breakdown of carbohydrates, fats, and amino acids to generate energy, are common to all living cells. Cancer cells, however, are characteristically different from surrounding healthy cells; in order to maintain viability and proliferate in frequently nutrient and oxygen-poor environments, cancer cells reprogram their metabolic activities [321 322 674]. The resulting shifts in intracellular and extracellular metabolite concentrations can have profound effects on the tumour microenvironment, cellular signalling, and gene expression. For this reason, metabolomics, the study of all the small molecular weight metabolites that make up the metabolome, is considered complementary to other, more established omic technologies.

The application of metabolomics has led to the identification of biomarkers for diagnosis and prognosis in multiple cancers, including prostate, brain and ovarian cancer [674-676]. My scoping review of the literature, however (Chapter 3), found that few studies have evaluated the metabolomic profiles of people with HNC, and even fewer studies have determined the

prognostic significance of circulating metabolites in this population. Those studies that have been conducted have included small sample sizes- typically fewer than 100 people, have generally considered multiple tumour sites together, and have used different types of samples (i.e., serum, plasma, saliva, urine, tissue) and analytical platforms (e.g. MS, NMR, liquid chromatography) [271]. For an overview of the different metabolomics technologies, the reader is directed to Chapter 4. One recent study evaluated the serum amino acid profiles of 140 people with histologically verified HNSCC using high-performance liquid chromatography in an attempt to identify possible prognostic biomarkers for OS and RFS [344]. The authors found serum methionine levels were positively correlated with five-year OS and RFS in multivariable models (HR, 0.52; 95% CI, 0.31–0.90, p = 0.02), which adjusted for age, sex, tumour localization, T stage (T3,4 vs. T1,2), N stage (N2,3 vs. N0,1), M stage (M1 vs. M0), histological tumour grade and treatment strategy. These findings have yet to be validated externally.

This chapter starts to explore the association of metabolite concentrations with all-cause mortality in a specific subgroup of people with HNC, namely those with oropharyngeal tumours, using NMR spectroscopy. Whilst exploratory in nature, this analysis is based on a larger dataset, with data available on a broad range of metabolites involved in multiple biological pathways, including amino acid metabolism, lipid metabolism and fluid balance. The findings of this study represent a first step towards identifying a metabolic signature, or profile, that could predict survival in people with OPC.

10.2. Aims and objectives

This analysis is not driven by an *a priori* hypothesis but is instead designed to be hypothesis-generating. The objectives are, firstly, to examine whether the metabolome of people with HPV-driven oropharyngeal tumours differs from that of people with non-HPV-driven oropharyngeal tumours, and secondly, to investigate possible associations between baseline metabolic traits and all-cause mortality in people with OPC using Cox proportional hazards regression models.

10.3. Methods

10.3.1. Study population

All participants with OPC and baseline blood samples were eligible for inclusion in the study. Samples were initially picked for metabolomic analysis using the information provided in the baseline data capture forms, that is, samples were selected based on clinical ICD-10 coding of OPC (see Chapter 3 for details). A minimum of 100 µL of blood was needed for the low volume NMR method, as described in Chapter 5. Samples were not picked if they were the last aliquot available for that participant. In total, 1,595 of 1,611 potentially eligible samples were available for analysis: 11 participants did not have enough sample volume available and 5 participants had no sample at all. Of these, 112 participants were excluded from the analysis because subsequent pathological reports revealed the cancer originated outside of the oropharynx. This left a final sample size of 1,483.

10.3.2. Measurement of metabolites

Metabolic profiling was performed on baseline plasma samples using a high-throughput serum NMR metabolomics platform (Nightingale Health®, Helsinki, Finland), originally described by Soininen et al. [590]. Details of this platform and its use in epidemiological studies have been described in Chapter 5 (*The Head and Neck 5000 study*). Briefly, the platform provides quantification in molar concentration units of: routine lipids; 14 lipoprotein subclasses, including particle concentration and lipids transported by these particles; various fatty acids and fatty acids traits (e.g., chain length, degree of unsaturation); amino acids; ketone bodies; glycolysis and gluconeogenesis-related metabolites; fluid balance; and one inflammation-related metabolite (Appendix D1). In total, 224 biochemically and metabolically distinct measures were obtained but for the purposes of analysis and for ease of interpretation, ratios (with the exception of fatty acid ratios), diameters, glycolysis-related metabolites (glucose, lactate and citrate) and alanine were excluded, leaving 145 discrete metabolic traits (Table 39). The rationale for retaining individual fatty-acid concentrations relative to total fatty acids is that previous work has shown that these measures reflect the biology of individual fatty acids better than their absolute concentration and they are commonly the only metric captured by standard laboratory assays [594]. Moreover, fatty acid ratios have been linked to cancer risk and progression [677 678]. The remaining selected metabolic measures were excluded on the basis of recommendations provided in the Nightingale Health report [679], which states:

"*Many metabolite concentrations differ from what we commonly observe. Many of these differences point to sample degradation. For example, the observed shifts in amino acids, (e.g. alanine) look similar than in samples that are kept for some time as whole blood (likely in room temperature) before plasma separation. In addition, the samples might have been stored at -20°C for extended periods of time….*

We note that most samples display very low glucose and elevated lactate levels. This is also commonly seen in the case when glycolysis has been ongoing after the sample collection (i.e. samples have been with the cells for some time prior to plasma separation). In this case, the concentrations no longer reflect the biological state of the study subject and thus we advise to exclude glycolysis related markers from the analyses"

This finding in H&N5000 can be explained by the fact that, following blood draws, samples were initially stored and posted at room temperature. Work by Ferreira *et al* demonstrates that pre-storage delay and incubation temperature of uncentrifuged samples can result in changes in levels of metabolite measures, particularly for glycolysis-related metabolites [680]. This is because the blood cells (primarily red blood cells) and enzymes contained within the sample tube are still metabolically active, resulting in the uptake and release of metabolites. By contrast, the same authors found that lipids, lipoproteins and fatty acids, were minimally affected by different sample pre-analytical conditions.

Overall, the mean success rate for metabolic trait detection in this analysis, which is the percentage of samples for which the respective metabolic trait measure was obtained, was 97.95% (included metabolites only; Table 43). Lipoprotein concentration measures were obtained for all samples. The largest amount of missing data was for glutamine (77% success rate). This is probably because glutamine is not very stable when the sample is stored at room temperature for prolonged periods. It degrades to glutamate. In total, 343 out of 1,595 samples were tagged "Low glutamine/High glutamate" in the Nightingale report. Glutamine was not quantified in these samples.

Table 43: An overview of the metabolic biomarkers included in the current analysis (n=145).

Abbreviation	Metabolite trait (unit)	Success rate*
XXL-VLDL-P	Concentration of chylomicrons and extremely large VLDL particles (mol/l)	100%
XXL-VLDL-L	Total lipids in chylomicrons and extremely large VLDL (mmol/l)	100%
XXL-VLDL-PL	Phospholipids in chylomicrons and extremely large VLDL (mmol/l)	100%
XXL-VLDL-C	Total cholesterol in chylomicrons and extremely large VLDL (mmol/l)	100%
XXL-VLDL-CE	Cholesterol esters in chylomicrons and extremely large VLDL (mmol/l)	100%
XXL-VLDL-FC	Free cholesterol in chylomicrons and extremely large VLDL (mmol/l)	100%
XXL-VLDL-TG	Triglycerides in chylomicrons and extremely large VLDL (mmol/l)	100%
XL-VLDL-P	Concentration of very large VLDL particles (mol/l)	100%
XL-VLDL-L	Total lipids in very large VLDL (mmol/l)	100%
XL-VLDL-PL	Phospholipids in very large VLDL (mmol/l)	100%
XL-VLDL-C	Total cholesterol in very large VLDL (mmol/l)	100%
XL-VLDL-CE	Cholesterol esters in very large VLDL (mmol/l)	100%
XL-VLDL-FC	Free cholesterol in very large VLDL (mmol/l)	100%
XL-VLDL-TG	Triglycerides in very large VLDL (mmol/l)	100%
L-VLDL-P	Concentration of large VLDL particles (mol/l)	100%
L-VLDL-L	Total lipids in large VLDL (mmol/l)	100%
L-VLDL-PL	Phospholipids in large VLDL (mmol/l)	100%
L-VLDL-C	Total cholesterol in large VLDL (mmol/l)	100%
L-VLDL-CE	Cholesterol esters in large VLDL (mmol/l)	100%
L-VLDL-FC	Free cholesterol in large VLDL (mmol/l)	100%
L-VLDL-TG	Triglycerides in large VLDL (mmol/l)	100%
M-VLDL-P	Concentration of medium VLDL particles (mol/l)	100%
M-VLDL-L	Total lipids in medium VLDL (mmol/l)	100%
M-VLDL-PL	Phospholipids in medium VLDL (mmol/l)	100%
M-VLDL-C	Total cholesterol in medium VLDL (mmol/l)	100%
M-VLDL-CE	Cholesterol esters in medium VLDL (mmol/l)	100%
M-VLDL-FC	Free cholesterol in medium VLDL (mmol/l)	100%
M-VLDL-TG	Triglycerides in medium VLDL (mmol/l)	100%
S-VLDL-P	Concentration of small VLDL particles (mol/l)	100%
S-VLDL-L	Total lipids in small VLDL (mmol/l)	100%
S-VLDL-PL	Phospholipids in small VLDL (mmol/l)	100%
S-VLDL-C	Total cholesterol in small VLDL (mmol/l)	100%
S-VLDL-CE	Cholesterol esters in small VLDL (mmol/l)	100%
S-VLDL-FC	Free cholesterol in small VLDL (mmol/l)	100%
S-VLDL-TG	Triglycerides in small VLDL (mmol/l)	100%
XS-VLDL-P	Concentration of very small VLDL particles (mol/l)	100%
XS-VLDL-L	Total lipids in very small VLDL (mmol/l)	100%
XS-VLDL-PL	Phospholipids in very small VLDL (mmol/l)	100%
XS-VLDL-C	Total cholesterol in very small VLDL (mmol/l)	100%
XS-VLDL-CE	Cholesterol esters in very small VLDL (mmol/l)	100%
XS-VLDL-FC	Free cholesterol in very small VLDL (mmol/l)	100%

XS-VLDL-TG	Triglycerides in very small VLDL (mmol/l)	100%

Table 43 continued.

Abbreviation	Metabolite trait	Success rate
IDL-P	Concentration of IDL particles (mol/l)	100%
IDL-L	Total lipids in IDL (mmol/l)	100%
IDL-PL	Phospholipids in IDL (mmol/l)	100%
IDL-C	Total cholesterol in IDL (mmol/l)	100%
IDL-CE	Cholesterol esters in IDL (mmol/l)	100%
IDL-FC	Free cholesterol in IDL (mmol/l)	100%
IDL-TG	Triglycerides in IDL (mmol/l)	100%
L-LDL-P	Concentration of large LDL particles (mol/l)	100%
L-LDL-L	Total lipids in large LDL (mmol/l)	100%
L-LDL-PL	Phospholipids in large LDL (mmol/l)	100%
L-LDL-C	Total cholesterol in large LDL (mmol/l)	100%
L-LDL-CE	Cholesterol esters in large LDL (mmol/l)	100%
L-LDL-FC	Free cholesterol in large LDL (mmol/l)	100%
L-LDL-TG	Triglycerides in large LDL (mmol/l)	100%
M-LDL-P	Concentration of medium LDL particles (mol/l)	100%
M-LDL-L	Total lipids in medium LDL (mmol/l)	100%
M-LDL-PL	Phospholipids in medium LDL (mmol/l)	100%
M-LDL-C	Total cholesterol in medium LDL (mmol/l)	100%
M-LDL-CE	Cholesterol esters in medium LDL (mmol/l)	100%
M-LDL-FC	Free cholesterol in medium LDL (mmol/l)	100%
M-LDL-TG	Triglycerides in medium LDL (mmol/l)	100%
S-LDL-P	Concentration of small LDL particles (mol/l)	100%
S-LDL-L	Total lipids in small LDL (mmol/l)	100%
S-LDL-PL	Phospholipids in small LDL (mmol/l)	100%
S-LDL-C	Total cholesterol in small LDL (mmol/l)	100%
S-LDL-CE	Cholesterol esters in small LDL (mmol/l)	100%
S-LDL-FC	Free cholesterol in small LDL (mmol/l)	100%
S-LDL-TG	Triglycerides in small LDL (mmol/l)	100%
XL-HDL-P	Concentration of very large HDL particles (mol/l)	100%
XL-HDL-L	Total lipids in very large HDL (mmol/l)	100%
XL-HDL-PL	Phospholipids in very large HDL (mmol/l)	100%
XL-HDL-C	Total cholesterol in very large HDL (mmol/l)	100%
XL-HDL-CE	Cholesterol esters in very large HDL (mmol/l)	100%
XL-HDL-FC	Free cholesterol in very large HDL (mmol/l)	100%
XL-HDL-TG	Triglycerides in very large HDL (mmol/l)	100%
L-HDL-P	Concentration of large HDL particles (mol/l)	100%
L-HDL-L	Total lipids in large HDL (mmol/l)	100%
L-HDL-PL	Phospholipids in large HDL (mmol/l)	100%
L-HDL-C	Total cholesterol in large HDL (mmol/l)	100%
L-HDL-CE	Cholesterol esters in large HDL (mmol/l)	100%
L-HDL-FC	Free cholesterol in large HDL (mmol/l)	100%
L-HDL-TG	Triglycerides in large HDL (mmol/l)	100%
M-HDL-P	Concentration of medium HDL particles (mol/l)	100%

M-HDL-L	Total lipids in medium HDL (mmol/l)	100%

Table 43 continued.

Abbreviation	Metabolite trait	Success rate
M-HDL-PL	Phospholipids in medium HDL (mmol/l)	100%
M-HDL-C	Total cholesterol in medium HDL (mmol/l)	100%
M-HDL-CE	Cholesterol esters in medium HDL (mmol/l)	100%
M-HDL-FC	Free cholesterol in medium HDL (mmol/l)	100%
M-HDL-TG	Triglycerides in medium HDL (mmol/l)	100%
S-HDL-P	Concentration of small HDL particles (mol/l)	100%
S-HDL-L	Total lipids in small HDL (mmol/l)	100%
S-HDL-PL	Phospholipids in small HDL (mmol/l)	100%
S-HDL-C	Total cholesterol in small HDL (mmol/l)	100%
S-HDL-CE	Cholesterol esters in small HDL (mmol/l)	100%
S-HDL-FC	Free cholesterol in small HDL (mmol/l)	100%
S-HDL-TG	Triglycerides in small HDL (mmol/l)	100%
Serum-C	Serum-C Serum total cholesterol (mmol/l)	100%
VLDL-C	Total cholesterol in VLDL (mmol/l)	100%
Remnant-C	Remnant cholesterol (non-HDL, non-LDL -cholesterol) (mmol/l)	100%
LDL-C	Total cholesterol in LDL (mmol/l)	100%
HDL-C	Total cholesterol in HDL (mmol/l)	100%
HDL2-C	Total cholesterol in HDL2 (mmol/l)	100%
HDL3-C	Total cholesterol in HDL3 (mmol/l)	100%
EstC	Esterified cholesterol (mmol/l)	98.6%
FreeC	Free cholesterol (mmol/l)	98.6%
Serum-TG	Serum-TG Serum total triglycerides (mmol/l)	100%
VLDL-TG	Triglycerides in VLDL (mmol/l)	100%
LDL-TG	Triglycerides in LDL (mmol/l)	100%
HDL-TG	Triglycerides in HDL (mmol/l)	100%
TotPG	Total phosphoglycerides (mmol/l)	98.6%
PC	Phosphatidylcholine and other cholines (mmol/l)	98.6%
SM	Sphingomyelins (mmol/l)	98.6%
TotCho	Total cholines (mmol/l)	98.6%
ApoA1	ApoA1 Apolipoprotein A (g/l)	100%
ApoB	Apolipoprotein B (g/l)	100%
TotFA	Total fatty acids (mmol/l)	96.6%
UnSat	Estimated degree of unsaturation	96.6%
DHA	Docosahexaenoic acid (mmol/l)	96.6%
LA	linoleic acid (mmol/l)	96.6%
FAw3	Omega-3 fatty acids (mmol/l)	96.6%
FAw6	Omega-6 fatty acids (mmol/l)	96.6%
pufa	PUFA Polyunsaturated fatty acids (mmol/l)	96.6%
mufa	Monounsaturated fatty acids (mmol/l)	96.6%
SFA	Saturated fatty acids (mmol/l)	96.6%
DHA/FA	Ratio of docosahexaenoic acid to total fatty acids (%)	96.6%
LA/FA	Ratio of linoleic acid to total fatty acids (%)	96.6%
FAw3/FA	Ratio of omega-3 fatty acids to total fatty acids (%)	96.6%

FAw6/FA	Ratio of omega-6 fatty acids to total fatty acids (%)	96.6%

Table 43 continued.

Abbreviation	Metabolite trait	Success rate
MUFA/FA	Ratio of monounsaturated fatty acids to total fatty acids (%)	96.6%
PUFA/FA	Ratio of polyunsaturated fatty acids to total fatty acids (%)	96.6%
SFA/FA	Ratio of saturated fatty acids to total fatty acids (%)	96.6%
Gln	Glutamine	77.4%
His	Histidine	99.9%
Ile	Isoleucine	99.6%
Leu	Leucine	99.6%
Val	Valine	99.7%
Phe	Phenylalanine	99.5%
Tyr	Tyrosine	99.4%
Ace	Acetate	98.2%
bOHBut	3-hydroxybutyrate	97.3%
Crea	Creatinine	97.3%
Alb	Albumin	100%
Gp	Glycoprotein acetylation	99.9%

** Success rate refers to the percentage of samples for which the respective metabolic trait measure was obtained.*

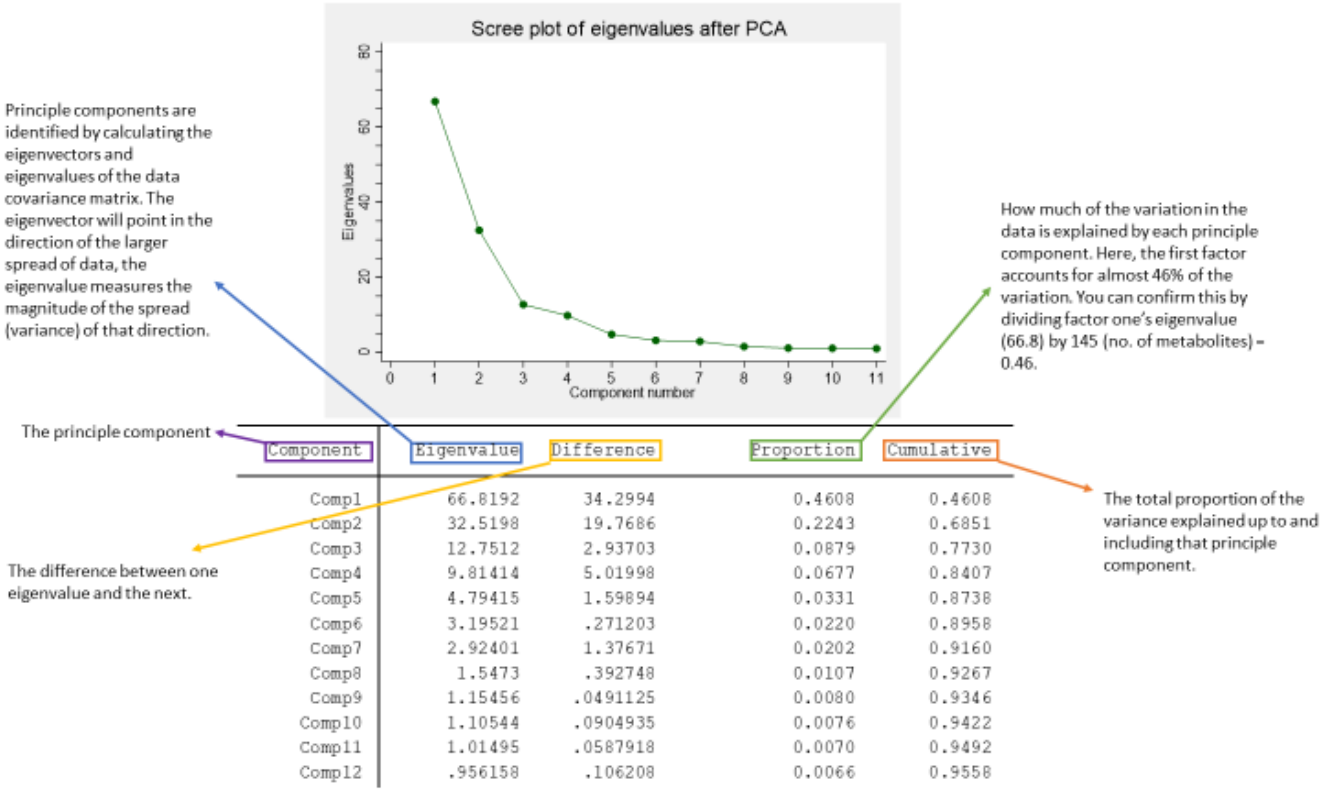
10.3.3. Issues of multiple testing

An important distinction between single biomarker studies and metabolomics analyses is the number of hypotheses being tested. Given that many metabolic measures are being analysed simultaneously, the probability of finding evidence of association by chance (i.e. type I errors, also known as false positives) is high. There are several methods available to help reduce the rate of type I errors [681 682]. In my analysis, I conducted principal component analyses (PCA) on standardised metabolic traits data. PCA is a data reduction technique that uses an orthogonal transformation, which is a type of linear transformation, to convert a series of observations of potentially correlated variables into a (smaller) set of uncorrelated variables, termed principal components (PCs) [683]. The first PC explains the maximum amount of the variability in the data and each subsequent PC explains as much of the remaining variability as possible. In this way, PCA reduces the dimensionality of a large dataset whilst preserving most of the information it contains. I used the number of PCs as the denominator in a Bonferroni correction to set a significance threshold which takes into account both multiple testing and the correlated nature of the metabolic traits, as discussed previously [601 684]. This method assumes that the independence of the principle components

(PCs) is equivalent to the degree of freedom of the original metabolic traits data and that retaining a number of PCs that is enough to explain at least 95% of the variance will result in only a small chance of type 1 error. In this instance, 95% of the variance in the 145 metabolic measures was explained by 11 PCs (this number is a proxy for the number of independent tests being performed). Therefore, the multiple testing correction, accounting for 11 independent tests using the Bonferroni method, resulted in $p < 0.005$ ($\alpha \div 11$, where $\alpha = 0.05$). *P*-values below this can be interpreted as providing strong evidence of an association of the respective metabolic trait with overall survival. A scree plot displaying the amount of variation (captured by the eigenvalues), that each principal component captures from the data is provided in Figure 50, alongside the PCA output obtained from Stata. The output shows that the first two PCs capture almost 69% of the overall variance in the data (PC1, 46%; PC2, 22%).

PC loadings were examined to determine which metabolite traits had the greatest influence on each component. Loadings represent the correlation between the original variables and the factors. Loadings range from -1 to 1, with loadings close to -1 or 1 indicating a strong influence on the component. Loadings close to zero, by contrast, show that the variable (trait) has a weak influence on the component. Considering that the first two PCs explain the majority of the variation in the data, I have focused primarily on their loadings. A table of loadings of all variables for each of the principal components that was studied is provided in Appendix D2. VLDL and LDL measures loaded highly on the first PC (PC1), alongside Apolipoprotein B (ApoB), remnant cholesterol (Remnant_C), monounsaturated fatty acids (mufa) and total fatty acids (TotFA) (Figure 51). HDL measures, particularly measures of total cholesterol in HDLs and ApoA1 were found to contribute considerably to the second PC (PC2). Of note, PC11 almost exclusively captures acetate, which has a loading of 0.86 (Appendix D3). The loadings of most of the other metabolites in PC11, with the exception of a few weakly influential metabolites (e.g. glutamine, histidine and creatinine) are close to zero.

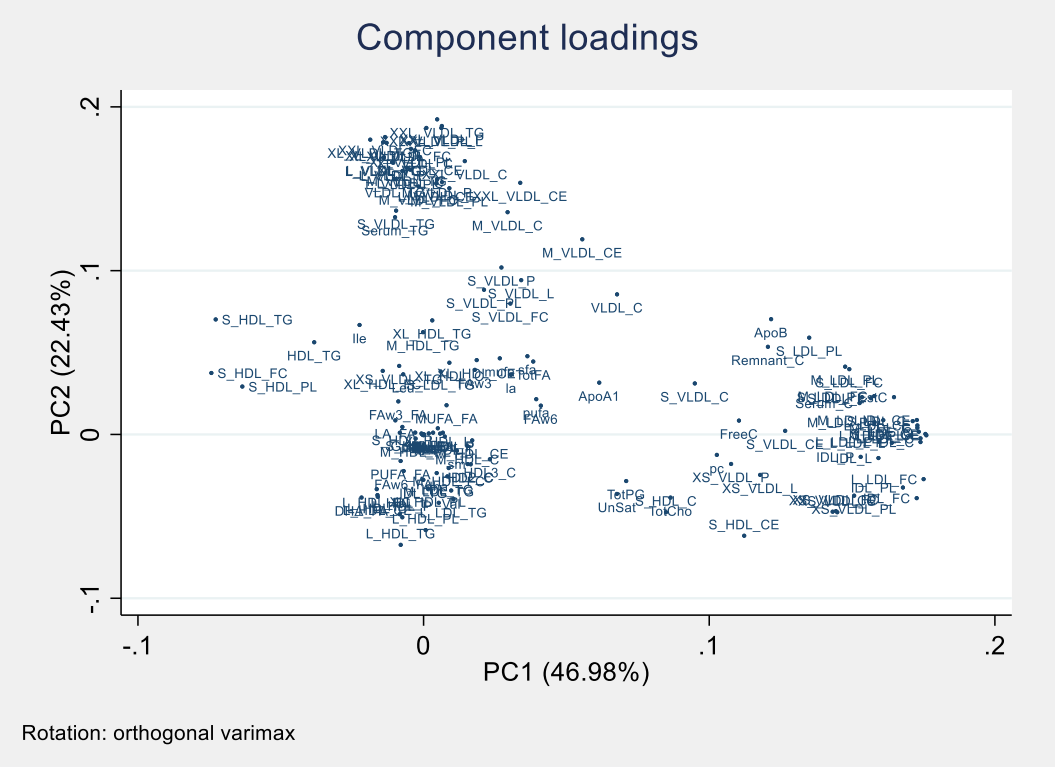
Figure 50: Scree plot and Stata output showing the decreasing rate at which variance is explained by additional principal components.



Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	66.8192	34.2994	0.4608	0.4608
Comp2	32.5198	19.7686	0.2243	0.6851
Comp3	12.7512	2.93703	0.0879	0.7730
Comp4	9.81414	5.01998	0.0677	0.8407
Comp5	4.79415	1.59894	0.0331	0.8738
Comp6	3.19521	.271203	0.0220	0.8958
Comp7	2.92401	1.37671	0.0202	0.9160
Comp8	1.5473	.392748	0.0107	0.9267
Comp9	1.15456	.0491125	0.0080	0.9346
Comp10	1.10544	.0904935	0.0076	0.9422
Comp11	1.01495	.0587918	0.0070	0.9492
Comp12	.956158	.106208	0.0066	0.9558

The scree plot (top) shows the eigenvalues on the y-axis and the number of factors on the x-axis. In these results, the first 11 principal components have eigenvalues greater than 1 and together explain almost 95% of the variation in the data (bottom). Kaiser criterion suggests retaining those factors with eigenvalues equal or higher than 1 [685].

Figure 51: Loading to PC1 and PC2.



Loadings have been rotated to maximise ease of interpretation using varimax rotation, which maximizes the sum of the variance of the squared loadings [686]. This typically produces high or low (near zero) factor loadings, with few intermediates, such that a variable is more likely to be associated with just one PC. The variance accounted for each PC is shown in brackets. VLDL and HDL measures load highly to PC1, whilst HDL measures contribute most to PC2.

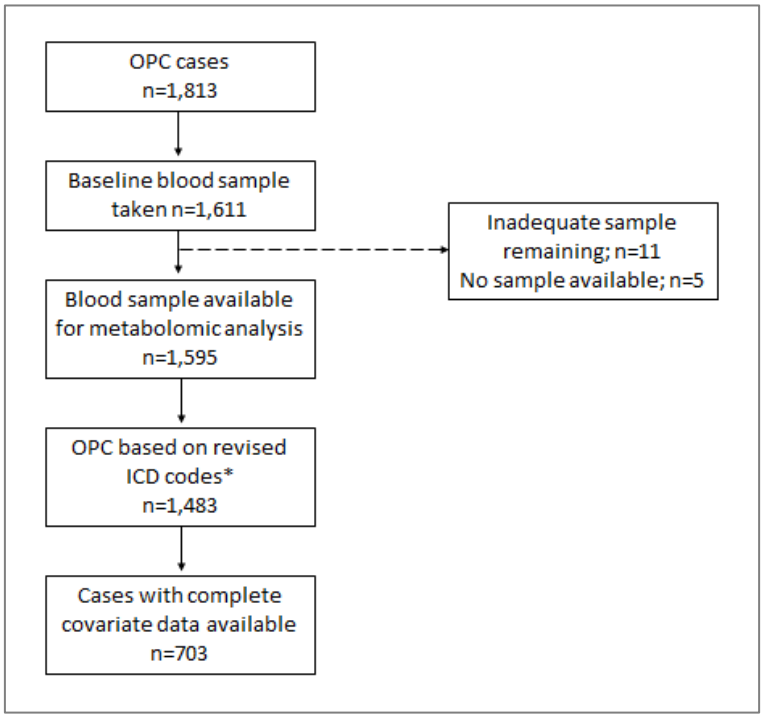
10.3.3. Missing data decisions

Missing covariate data were not imputed in the current analysis. This was because the imputation model would necessarily have included all the metabolic trait measures, since all of the variables included in the substantive model should be included in the imputation model. This was problematic for two reasons. Firstly, imputation by multiple chained equations (ICE) assumes normality for all of the continuous variables in the imputation model [687] but several of the metabolic traits exhibited a skewed distribution (see Appendix D4). This issue could not be corrected using log, square root, or reciprocal transformations. Indeed, several of the normally distributed variables became skewed after transformation.

Several nonnormal imputation models are currently in development [688], but this remains an active area of research. Secondly, as mentioned previously, ICE assumes that missing values are missing completely at random (MCAR) or missing at random (MAR). This is an issue because there was not enough information to determine why metabolite data were missing. Missingness among metabolite data is a common occurrence in epidemiological studies [681], though it is generally less of an issue with NMR data compared to MS data (personal communication), in part because NMR requires limited sample preparation (e.g. separation or derivatization) and also because of its high selectivity [689]. Values may be missing for several reasons including biological factors (e.g. metabolites being absent), technical limitations (e.g. the limit of detection or issues separating metabolite signals from noise), or measurement error [690].

In light of these two issues, it was decided that imputation would not be appropriate. For information on the proportion of missing covariate data, see Appendix D3. This left 703 of 1813 (39%) participants with OPC for inclusion in the analysis (Figure 52).

Figure 52: OPC samples available for analysis



Abbreviations: ***ICD****, International classification of diseases;* ***n****, number;* ***OPC****, oropharyngeal cancer.* ∗ *Pathologically confirmed where possible.*

10.3.4. Statistical analyses

Analyses were undertaken in the statistical software packages RStudio (version 1.2.5001) and StataMP (version 15.1).

To allow comparison of magnitudes of association across measures with different units or with large differences in their concentration distributions, all metabolite concentrations were converted to standard deviation (z) scores.

First, to examine whether there were differences in the baseline metabolomes of HPV-positive and HPV-negative individuals (based on HPV16-E6 serological status, which is a good marker of HPV-driven tumours [628]), metabolic traits were regressed on HPV status to get the difference in mean values by HPV status with 95% CIs and p-values for the differences. Regression coefficients (Betas) represent the SD difference in mean metabolite concentration for HPV-positive versus HPV negative individuals. Associations between OS and metabolite measures (in all OPCs) were then estimated as HRs and 95% CIs for each metabolite trait using the Cox proportional hazards model. The PH assumption was tested using statistical tests based on the Schoenfeld residuals. There was no evidence for non-proportionality for any of the metabolic measures or covariates included in the models.

Four separate models were run on the 703 included participants:

1) a minimally adjusted model that adjusted for age and gender;
2) a model that additionally adjusted for clinical factors including TNM stage, HPV status, comorbidity and BMI;
3) a model that additionally adjusted for socioeconomic factors, namely annual household income, educational attainment and marital status; and
4) a model that additionally adjusted for smoking status and alcohol intake.

Given that complete case analyses can lead to biased estimates if the associations amongst individuals included in the analysis differ from the population from which they are drawn, minimally adjusted models were run in the full dataset and in the complete case dataset and the coefficients were compared. The correlation between datasets was assessed using the R_2 measure.

10.3.5. Sensitivity analysis

The observed concentrations and distributions of metabolite measures are presented graphically in Appendix D4. Given that several of the metabolic traits exhibit a skewed distribution, I examined the effect of removing potential outliers, i.e. samples which deviate from the distribution of the majority of the data, from the dataset. Inclusion or exclusion of these outliers could be important and may lead to different statistical conclusions, particularly given the relatively modest sample size of this study. I performed three different sensitivity analyses: 1) I removed values that were greater than 5 SD from the mean; 2) I removed values above the 99th percentile for that variable; 3) I winsorized the top 5% and bottom 5% of data points.

The SD approach for defining outliers, whereby values are classified as outliers if they are a given number of SDs away from the mean score for that variable- typically 3 SD, is a commonly applied rule [691]. The practice of using plus or minus 3 SD is based on the characteristics of a normal distribution, where 99.7% of the data falls within this range [692]. Some metabolomics studies use a cut-off of 4 or 5 SD from the mean [693] [694]. I decided to remove values that were greater than 5 SD from the mean, partly to preserve sample size and thus statistical power but also because human blood metabolites are known to vary widely across different individuals [695]. For comparison, I used the slightly less conservative approach of defining outliers as values above the 99th percentile for that variable.

The SD rule has been criticised as an approach for removing outliers for the reason that the mean and SD are very sensitive to extreme values i.e. outliers increase the SD [696] (though using a larger SD cut-off mitigates this to some extent). Moreover, the method assumes a normal distribution. For these reasons, this approach may be fundamentally problematic in the context of metabolomics data, where, as is the case in this analysis, data typically do not conform to normal distributions. Another approach which has been applied in metabolomic studies is winsorization [697] [698] [699], whereby extreme data values i.e. values in the tails of the distribution, are replaced with smaller values. In this way, winsorization is not equivalent to removing data, but rather data that is assumed to be incorrect or exaggerated is replaced with a more plausible value. The new value therefore represents a compromise.

As a final sensitivity analysis, all metabolite variables were winsorized using values representing the 5th and 95th percentiles of the distribution in the sample (i.e. data points lying outside the 5th percentile were replaced with the 5th percentile for that metabolite

distribution and data points lying outside the 95th percentile were replaced with the 95th percentile). This was implemented in Stata using the `winsor2` command.

10.4. Results

10.4.1. Baseline characteristics

The primary survival analysis was performed in 703 individuals, of whom 129 died during a median follow-up time from diagnosis of 4.3 years (IQR: 3.8, 5.1). Over two-thirds of the population were HPV-positive and the majority were diagnosed with stage IV tumours. A full description of the baseline demographics of the sample population was provided in Chapter 6 (overall, Table 23; stratified by HPV status, Table 24).

When the baseline clinical and socio-demographic variables for participants included in the analytic sample were compared against those of the people who were excluded from the analysis (i.e. those with missing covariate data), there was evidence that people with missing data were more likely to be HPV-negative (33%% vs. 26%; *p*-value for difference =0.004), were more likely to have severe comorbidity (20% vs. 14%; p<0.001), and were more likely to be current smokers (22% vs. 15%l; p=0.005) compared with people who had complete data (Table 44). The mean age of people included in the analytic sample was around a year older than that of the excluded group (45 years vs. 52 years for the complete case group compared to the group with missing data).

10.4.2. Differences in metabolic trait concentrations between HPV-positive and HPV-negative individuals

The SD difference in mean metabolic trait concentrations in HPV-positive versus HPV-negative individuals is presented in (Appendix D5). There was evidence of a difference in approximately 75% of the traits measured ($p \leq 0.005$). The strongest evidence of a difference was obtained for large HDL (L_HDL) measures, including cholesterol esters in L_HDL (β= -0.50, 95% CI: -0.67, -0.34; $p=2.1 \times 10^{-09}$), total cholesterol in L_HDL (β= -0.50, 95% CI: -0.66, -0.34; $p=2.55 \times 10^{-09}$), free cholesterol in L_HDL (β= -0.49, 95% CI: -0.66, -0.33; $p=4.99 \times 10^{-09}$), total lipids in L_HDL (β= -0.47, 95% CI: -0.63, -0.31; $p=2.52 \times 10^{-08}$) and

concentration of L_HDL particles (β= -0.47, 95% CI: -0.63, -0.30; p=3.12 x 10^{-08}). Other metabolite traits that showed a difference between HPV groups included, but were not restricted to, creatinine (β=0.41, 95% CI: 0.25, 0.56; p=1.77 x 10^{-07}), the amino acids valine (β=0.41, 95% CI: 0.25, 0.57; p=6.14 x 10^{-07}), histidine (β=0.42, 95% CI: 0.26, 0.59; p=8.86 x 10^{-07}), linoleic acid (β=0.40, 95% CI: 0.22, 0.57; p=8.87 x 10^{-06}), leucine (β=0.33, 95% CI: 0.16, 0.51; p=1.07 x 10^{-04}) and omega-3 fatty acids (β=0.33, 95% CI: 0.15, 0.51; p=2.73 x 10^{-04}).

10.4.3. Associations of pre-treatment metabolic traits with all-cause mortality

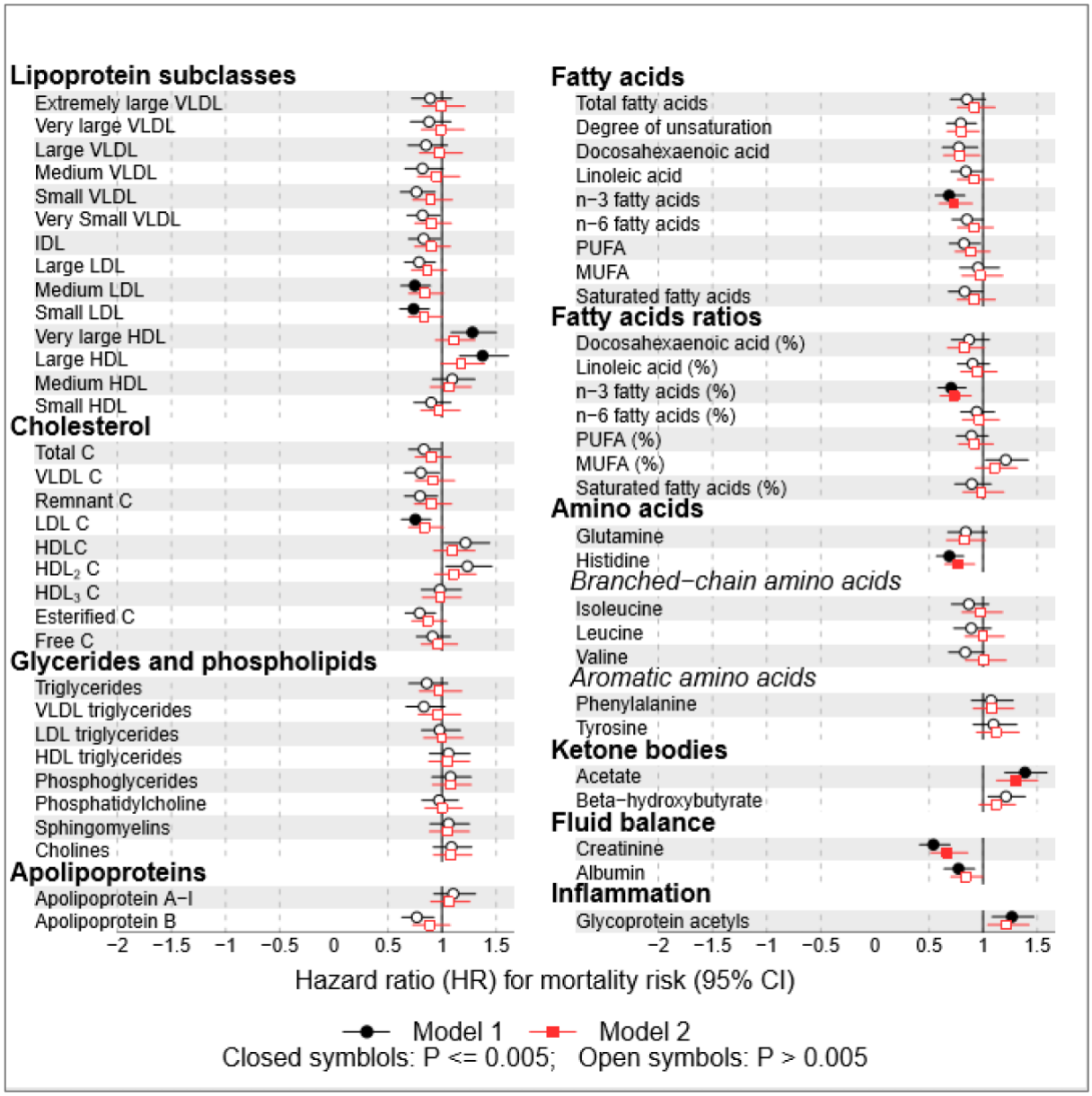
The results of the Cox regression analysis are presented in Figures 53 and 54. Based on the threshold for multiple testing (p=0.006), there was evidence of an association between 37 of the metabolic traits and survival in the minimally adjusted model (model 1). Of these, 29 were lipoprotein measures (see Appendix D6 for a complete list). After additionally controlling for clinical factors (model 2), five metabolic measures were related to OS (Figure 53). Each SD increase in acetate was associated with a 30% increased risk of death (HR=1.30; 95% CI: 1.12, 1.51; p=4.04 x 10^{-4}), whilst SD increases in creatinine, omega-3, the ratio of omega-3 to total fatty acids, and histidine were associated with reduced risk (creatinine, HR=0.67, 95% CI: 0.52, 0.86; p=0.002; omega-3, HR=0.73, 95% CI: 0.59, 0.90; p=0.003; ratio of omega-3 to total fatty acids, HR=0.73, 95% CI: 0.60, 0.89; p=0.002; histidine, HR= 0.77, 95% CI:0.65, 0.92; p=0.004). The associations of acetate and creatinine with OS remained in model 3 (acetate: HR=1.30, 95% CI: 1.11, 1.51; p=0.001; creatinine: HR=0.68; 95% CI: 0.53, 0.89; p=0.004), indicating that socioeconomic variables had little confounding effect on the observed associations (Figure 54). However, there was no longer good statistical evidence to suggest that omega 3 and histidine were related to mortality risk. Only acetate was associated with mortality risk in the fully adjusted model (model 4), which also adjusted for smoking status and alcohol intake (Figure 54). The effect estimate attenuated slightly (HR=1.28, 95% CI: 1.10, 1.49; p=0.002). An extended table of results can be found in Appendix D6.

Examining the linear correlation plot, comparing the minimally adjusted results for the complete case dataset and the full dataset (Figure 55), there appears to be good overall correlation between HRs, as indicated by an R_2 of 0.82 and a slope of 1.01 (± 0.04).

Table 44: A comparison of people included and excluded from the analytic sample.

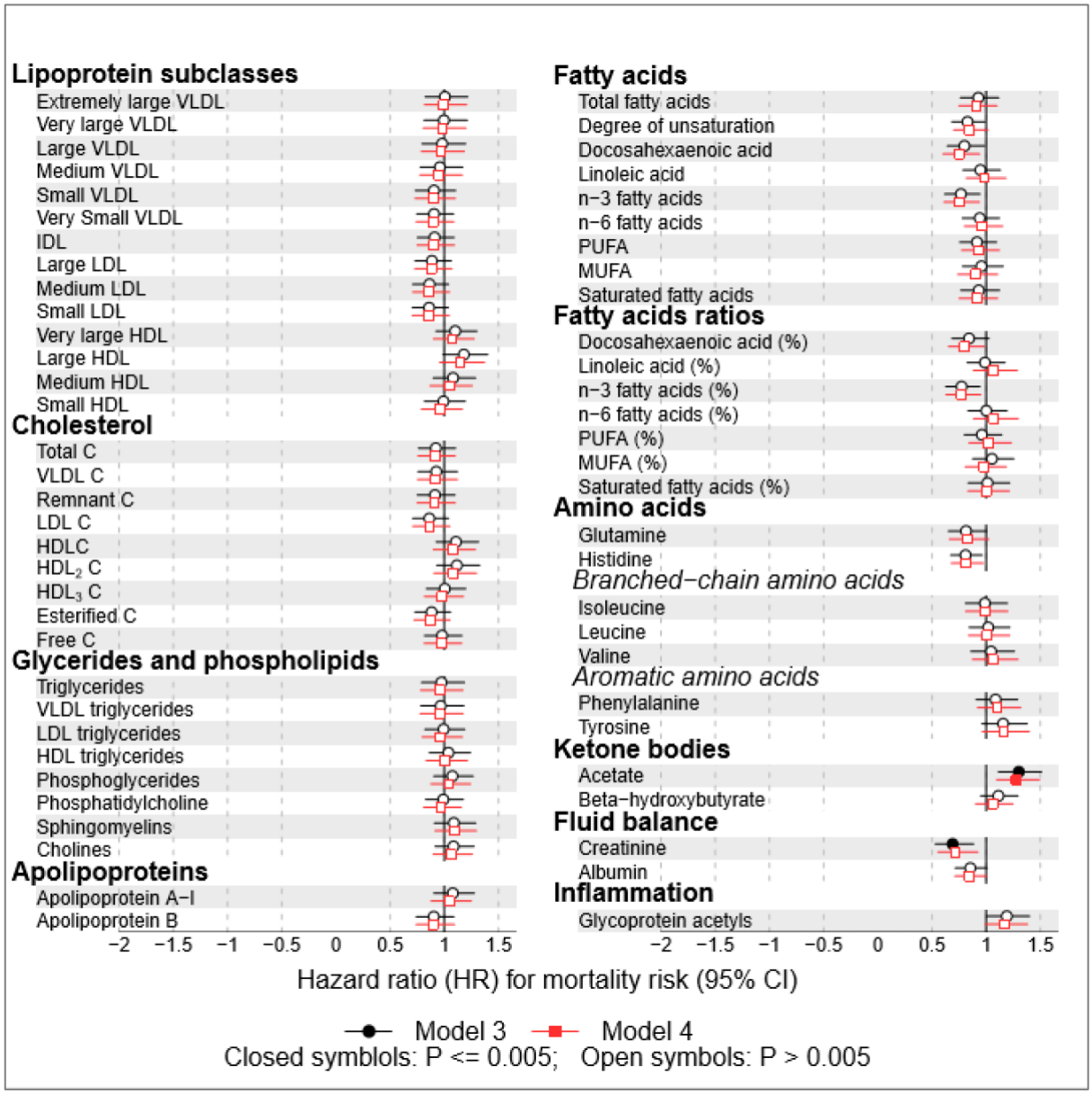
	Have missing data (n=780)		Complete case (n=703)		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*p-value*
Sex					
Male	599	76.80%	564	80.20%	
Female	181	23.20%	139	19.80%	0.109
TNM staging					
I	30	3.90%	24	3.40%	
II	70	9.00%	73	10.40%	
III	121	15.60%	94	13.40%	
IV	557	71.60%	512	72.80%	0.528
HPV serology group					
HPV-negative	258	33.20%	185	26.30%	
HPV-positive	518	66.80%	518	73.70%	0.004
Comorbidity groups					
None	321	42.50%	379	53.90%	
Mild	282	37.30%	223	31.70%	
Moderate/Severe	153	20.20%	101	14.40%	<0.001
Education level					
School education	176	45.60%	293	41.70%	
College	146	37.80%	272	38.70%	
Degree	64	16.60%	138	19.60%	0.335
Annual household income					
<£18,000	108	35.90%	249	35.40%	
£18000-£34,999	94	31.20%	222	31.60%	
>£35,000	99	32.90%	232	33.00%	0.989
Relationship status					
Single (never married)	56	13.30%	68	9.70%	
Currently in relationship	288	68.40%	519	73.80%	
No longer with spouse	77	18.30%	116	16.50%	0.095
Smoking status					
Never	95	24.90%	216	30.70%	
Former	201	52.80%	382	54.30%	
Current	85	22.30%	105	14.90%	0.005
Alcohol consumption					
Non-drinker	104	25.70%	179	25.50%	
Moderate drinker	80	19.80%	171	24.30%	
Hazardous-harmful drinker	221	54.60%	353	50.20%	0.191
	N	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	770	59.49 (9.11)	703	58.21 (9.07)	0.007
BMI	177	26.79 (5.27)	703	26.98 (5.05)	0.647

Figure 53: Cox regression results for models 1 and 2



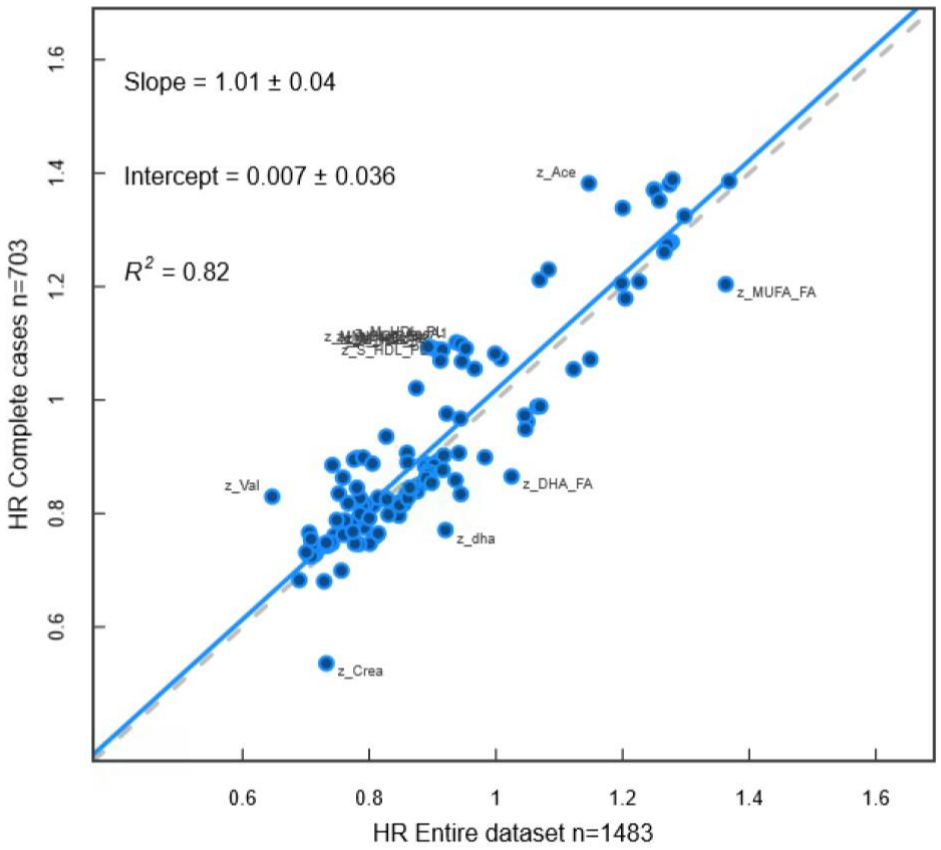
Hazard ratios represent the increase in mortality risk per SD increase in metabolic trait. Model 1 (minimally adjusted) adjusted for age and gender; model 2 (clinical model) additionally adjusted for TNM stage, HPV status, comorbidity and BMI. Lipoprotein metabolic trait measures have been grouped for the purposes of presentation.

Figure 54: Cox regression results for models 3 and 4.



Hazard ratios represent the increase in mortality risk per SD increase in metabolic trait. Model 3 (socioeconomic model) adjusted for income, education and marital status; model 4 (fully adjusted model) additionally adjusted for TNM stage, HPV status, comorbidity and BMI. Lipoprotein metabolic trait measures have been grouped for the purposes of presentation.

Figure 55: Linear fit between complete case and full dataset models.



Each blue dot represents a metabolic trait and the positions of the dots are determined by its association with survival. A linear fit of the overall correspondence summarises the similarity in effect estimates (blue solid line) between datasets. R_2 indicates goodness of linear fit and as such is a measure of the consistency between associations. A slope of 1 with an intercept of 0 (grey dashed line), with all blue dots sitting on the line (R_2 = 1), would indicate that the full and compete case dataset estimates have the same magnitude and direction. Metabolic traits whose residuals are within the top 10% are labelled. Only metabolic traits with the largest residuals (top 10%) are marked.

10.4.4. Sensitivity analyses

A comparison of the effect estimates obtained for selected metabolic traits in the primary and sensitivity analyses is presented in Table 45. Only those metabolites for which there was evidence of an association with survival in the primary analysis are included.

When data points (samples) that were in the top 99th percentile were removed from the data set and the analysis was repeated (n=515 remaining), the effect estimates for omega-3 and histidine in the minimally adjusted model were very similar to those of the primary analysis (HRs = 0.65 [95% CI: 0.51, 0.83; *p*=0.001] and 0.69 [0.55, 0.87; *p*=0.002], respectively), but the improvement in mortality risk associated with each SD increase in acetate decreased (HR= 1.1 [95% CI: 0.99, 1.26; *p*=0.085]) as did the higher risk associated with creatinine (HR = 0.62 [95% CI: 0.47, 0.82; p=0.001]. In the fully adjusted model, the effect estimates attenuated for acetate and omega 3, compared with the primary analysis (HR acetate = 1.13 [95% CI:0.99, 1.28; *p*=0.072]; HR omega-3 = 0.69 [95% C/I:0.53, 0.89; *p*=0.005]) but were broadly comparable for histidine (HR=0.84 [95% CI: 0.67, 1.06; p=0.142]). There HR for creatinine was slightly higher (HR=0.78 [95% CI: 0.60, 1.02; *p*=0.071), though CIs overlapped with those of the primary analysis.

Using the more conservative approach of removing observations within 5 SD from the mean (n=505 remaining), the results for creatinine, omega-3 and histidine in the minimally adjusted models were similar to those of analysis described above, so effect estimates for omega-3 and histidine were comparable with the primary analysis (HR=0.65 [95% CI: 0.50, 0.83; *p*=0.001] for omega-3 and 0.66 [95% CI:0.52, 0,83; *p*=0.001] for histidine) but the mortality risk associated with creatinine reduced (HR=0.62 [95% CI: 0.47, 0.82; *p*=0.001]; Table 43). There was no longer strong evidence that increased acetate improved OS (HR=0.85 [95% CI: 0.64, 1.14; *p*=0.280]). A similar pattern was seen in the fully adjusted models.

When metabolite variables were winsorized, which preserved the original sample size but replaced extreme values in the top and bottom 5% with the 95th and 5th percentiles for that metabolite distribution, the results for creatinine, omega-3 and histidine were again comparable to those if the primary analysis: minimally adjusted HRs: 0.59 [95% CI: 0.48, 0.72; *p*=4.56 x 10^{-04}] for creatinine, 0.69 [95% CI: 0.57, 0.83; *p*=1.29 x 10^{-04}] for omega-3 and 0.69 [95% CI: 0.57, 0.83; *p*=9.93 x 10^{-04}] for histidine; corresponding fully adjusted HRs= 0.75 [95% CI: 0.61, 0.93; *p*=0.008], 0.77 [95% CI: 0.62, 0.94; *p*=0.011], and 0.81 [95% CI:

0.67, 0.97; *p*=0.023]. Again, there was no real evidence to suggest that acetate influenced mortality risk in this analysis (minimally adjusted HR=0.99 [95% CI: 0.83, 1.18; *p*=0.910]; fully adjusted HR=0.99 [95% CI: 0.83, 1.17; *p*=0.884].

Table 45: The impact on effect estimates of removing potential outliers from the dataset, using different outlier-detection methods.

	Min adjusted					**Fully adjusted**				
			95% CI					95% CI		
	N	HR	Lower	Upper	*p*-value	N	HR	Lower	Upper	*p*-value
Primary analysis										
Acetate	696	1.38	1.20	1.59	<0.001	696	1.28	1.10	1.49	0.002
Creatinine	685	0.54	0.41	0.70	<0.001	685	0.72	0.55	0.92	0.010
Omega-3	680	0.68	0.56	0.83	1.66E-04	680	0.76	0.61	0.94	0.011
Histidine	703	0.68	0.57	0.82	5.34E-05	703	0.81	0.68	0.97	0.024
Remove top 99th percentile										
Acetate	515	1.11	0.99	1.26	0.085	515	1.13	0.99	1.28	0.072
Creatinine	515	0.62	0.47	0.82	0.001	515	0.78	0.60	1.02	0.071
Omega-3	515	0.65	0.51	0.83	0.001	515	0.69	0.53	0.89	0.005
Histidine	515	0.69	0.55	0.87	0.002	515	0.84	0.67	1.06	0.142
remove values 5 SD from mean										
Acetate	505	0.85	0.64	1.14	0.280	505	0.91	0.70	1.18	0.474
Creatinine	505	0.62	0.47	0.82	0.001	505	0.77	0.58	1.01	0.056
Omega-3	505	0.65	0.50	0.83	0.001	505	0.72	0.56	0.93	0.012
Histidine	505	0.66	0.52	0.83	0.001	505	0.80	0.63	1.01	0.061
Winsorized										
Acetate	696	0.99	0.83	1.18	0.910	696	0.99	0.83	1.17	0.884
Creatinine	685	0.59	0.48	0.72	4.56E-07	685	0.75	0.61	0.93	0.008
Omega-3	680	0.69	0.57	0.83	1.29E-04	680	0.77	0.62	0.94	0.011
Histidine	703	0.69	0.57	0.83	9.93E-05	703	0.81	0.67	0.97	0.023

10.5. Discussion

10.5.1. Principle findings

Changes in intra- and extracellular metabolism are well documented in the cancer literature, yet there are currently no established metabolic biomarkers for prognosis in OPC. Utilizing NMR plasma metabolomic profiling of 703 people with OPC, acetate, creatinine, omega-3 and histidine levels were found to be associated with OS in models that controlled for demographic (age and gender) and clinical (TNM stage, HPV status, comorbidity and BMI) factors. Each SD unit increase in circulating acetate was associated with a 30% higher risk of death (HR= 1.30, 95% CI: 1.12, 1.51; p=4.04 x 10^{-4}), whilst corresponding increases in circulating creatinine, omega-3 and histidine were associated with reduced mortality risks (creatinine HR= 0.67, 95% CI: 0.52, 0.86; p=0.002; omega-3 HR=0.73, 95% CI: 0.59, 0.90; p=0.003; histidine HR= 0.77, 95% CI: 0.65, 0.92; p=0.004). A higher ratio of omega-3 fatty acid relative to total fatty acid was also associated with improved survival in this analysis (HR= 0.73, 95% CI: 0.60, 0.89; p=0.002). The effect estimates for acetate and creatinine did not change considerably following adjustment for socioeconomic factors (income, education and marital status), however, evidence for the association of creatinine with survival was no longer present when models were adjusted for smoking and alcohol intake.

Whilst this study identified several potential biomarkers of interest, the findings should be interpreted with caution because the results of the sensitivity analyses suggest that some of the associations are driven by extreme values (outliers) or sample size. For example, higher acetate levels were associated with worse survival in the primary analysis but there was no evidence for this after extreme observations were removed. The effect estimates for creatinine, omega-3 and histidine were more stable across the analyses, but further, adequately powered studies are required to substantiate these findings.

The positive association between circulating acetate levels and all-cause mortality in this population of people with OPC, may reflect the observation that acetate provides an important respiratory substrate for metabolically stressed cancer cells [700-703]. In well-oxygenated (i.e. normoxic) conditions, acetyl coenzyme A (acetyl-CoA), a key precursor of lipid synthesis and energy production [704], is predominantly generated from glucose and glutamine-derived carbon. Under hypoxic conditions, however, isotopic tracer experiments have revealed that acetate, obtained from either external (e.g. diet) or internal sources, provides an alternative carbon source for biosynthesis, fuelling cellular proliferation [700 702 705].

Furthermore, it has been suggested that many cancer cells actually favour acetate as their carbon source [701 706 707].

Exogenous acetate can be produced from the oxidation of ethanol (alcohol) [702]. Typically, blood concentrations of ethanol range from 50–200 μM but among chronic drinkers, concentrations may reach 500–600 μM, or higher [702]. This is noteworthy given that alcohol is a known risk factor for HNC and many individuals with HNC have a history of chronic heavy alcohol consumption. It is possible that the observed association of acetate with all-cause mortality in this analysis may be driven by heavy drinkers, since heavy drinking is strongly associated with a range of harmful traits including poor dietary habits, lower quality of life and greater risk of depression [708], factors which may all negatively impact upon mortality. If acetate is on the causal pathway, i.e. if increased acetate is a consequence of alcohol consumption and part of the mechanism by which it increases the risk of death, adjusting for alcohol in the analysis (model 4) may not have been appropriate as this would suggest it is not a confounder.

The suggestion of a possible link between creatinine and mortality risk may warrant further investigation. Creatinine is a by-product of muscle metabolism, a waste product left behind when the body uses the amino acid creatine for energy [709 710]. In healthy individuals, creatinine is filtered from the blood by the kidneys but if an individuals' kidneys are not functioning properly, creatinine may accumulate in the blood. As such, serum creatine levels are widely used as a marker of renal function (glomerular filtration rate, GFR) in clinical practice. Causes of low blood creatinine concentration include low muscle mass (because the breakdown of muscle tissue produces creatinine), liver disease, fluid overload and poor nutritional status [710]. Reduced muscle mass - a condition known as sarcopenia - and malnourishment are common manifestations of HNC and cancer in general, and are frequently associated with poor clinical outcomes including reduced response to cancer treatment, treatment toxicity, and worse overall-survival [711-715]. In view of this, it is possible that any association of plasma creatinine levels with mortality may reflect the association of low muscle mass with mortality [710 716].

The finding that circulating omega-3 may be related to survival in this analysis was also interesting. As mentioned above, people with HNC are frequently malnourished and experience a high incidence of postoperative complications. It has been suggested that supplementation with omega-3 fatty acids could reduce the incidence of these complications, by maintaining immunocompetence during treatment [717-721]. It is difficult to draw any conclusions however, owing to the diversity of interventions used (i.e. formulations, durations of the

intervention, clinical outcomes studied), and the type and stage of cancers studied [721]. The finding that higher omega-3 levels were associated with improved survival in this analysis (in models that controlled for age, gender and clinical factors), provides further motivation for designing an adequately powered randomised controlled trial (RCT) investigating the effects of omega-supplementation on clinical outcomes in HNC.

10.5.2. Challenges

I encountered several challenges when trying to investigate the association of metabolites with survival, and this is a reflective of the relative infancy of the field compared with other omics technologies. Statistical methodologies for analysing genetic and epigenetic data are more developed and more widely discussed in the literature, for example compared with metabolomics, though methods and technologies are advancing rapidly. The first problem that I faced was how to deal with the skewed distribution of the metabolic data. There does not appear to be a standard procedure for this and despite trying several different approaches, I was unable to transform my data so that the distributions were all approximately normal, indeed some methods such as log transformation made some of the variables more skewed.

My second challenge was how to deal with the high dimensionality and inter-correlation of the metabolite data, which meant that the number of variables in my dataset greatly exceeded the number of samples and my burden of multiple testing- an issue which is pertinent to all 'omics' studies (genomics, epigenomics, transcriptomics and proteomics). At present, there are no existing standard protocols or optimal methodologies for analysing metabolomics data [722]. Some previous studies have used statistical methods such as Bonferroni correction and false discovery rate (FDR) to reduce the occurrence of false-positives in metabolomics studies [681], but in other fields of study such as genetics and epigenetics, i.e. where high-dimensional datasets are analysed, these corrections has been found to be too conservative [723]. This can lead to features of interest being missed. Here, a PCA approach, which been used in prior human metabolomics studies [680], was used to set a significance threshold which takes account of both the inter-correlation between individual metabolites and the number of statistical tests being performed. By taking account of the intercorrelation, this avoids generating over-conservative *p*-value thresholds.

A third issue that I had, related to the distribution of the metabolites, was that there were a number of possible outliers in my data. This is potentially problematic as the presence of extreme values can lead to spurious associations between an exposure(s) of interest and an

outcome(s). However, there is no clear consensus on how to deal with outliers in metabolomic studies, or indeed whether they should be 'dealt with' at all. Ideally, one would have knowledge of the range of values that is biologically plausible for all measures and then one could make an informed judgement as to whether an observed value is likely. Alternatively, one would have repeat samples for each individual, in which case an average value could be taken, for example. I examined the effect of removing outliers using different techniques to determine whether this impacted on the interpretation of my findings.

10.5.3. Strengths and limitations of the study

The main strengths of this analysis include its large sample size and the wide range of metabolic traits studied, facilitated through the use of high-throughput NMR, which, as described in Chapter 4, is highly reproducible [724]. I am not aware of any other HNC study with a similar (or larger) sample size and detailed prospectively collected phenotypic and metabolomic measurements. The metabolite traits captured by the targeted NMR platform are implicated in multiple biological pathways including lipoprotein and fatty acid metabolism, the citric acid cycle, amino acid metabolism and fluid balance. The platform itself has been used in numerous epidemiological and genetic studies, with over 200 publications having applied these methods to date [725].

Limitations should be considered when evaluating the results of this study. The first major limitation is that samples were initially stored and posted at room temperature, meaning that it was not possible to include glycolysis related measures in the analyses. Specifically, glucose, lactate and citrate, biomarkers which are available from EDTA plasma samples (glycine, pyruvate and glycerol are only available for serum samples [725]), could not be considered. Glycolysis provides the precursors for several major macromolecules including proteins, lipids, carbohydrates and nucleic acids which are all necessary to support cancer cell proliferation. Indeed, aerobic glycolysis or the 'Warburg effect' (rather than oxidative phosphorylation) is a key metabolic hallmark of nearly all cancers [322]. Future studies, aimed specifically at analysing the HNC metabolome, should take these pre-analytical factors into consideration in order to preserve the metabolites present at the time of sampling and achieve the most complete metabolic profile possible. Currently, there does not appear to be a standard operating procedure (SOP) for sample acquisition in metabolomic studies [722 726] and not many published papers provide a clear account of the methods they used. However, all studies should aim to process and freeze samples as quickly as possible. Ideally, samples should be centrifuged immediately (4°C) or briefly kept on ice if replicates are to be

collected (personal communication). They should be aliquoted to prevent freeze thaw cycles and stored at -80°C. Storing them at this temperature ensures that sample quality is maintained in the long term. If samples need to be shipped, they should be transported on dry ice. Crucially, there must be consistency in sample collection and handling in order to minimize variance/systematic bias, enabling meaningful results to be obtained.

A second but related issue is that only one sample was available for each participant. The metabolome is inherently dynamic, more so than the proteome or the genome since it provides a reflection of the biological processes or states at the time of sample collection [727]. This could be problematic and therefore, ideally, blood draws would have been repeated on the same individual at one or more time points to account for biological variability. Metabolon (an alternative NMR platform) state that repeated sampling from the same subject reduces the number of study subjects needed to achieve a sufficiently well-powered study [728]. Again, future studies should look to include multiple blood sampling in their study design. In addition, it would be interesting to look at the associations of metabolites with survival in other biological samples like saliva or urine.

A third potential limitation of the study is that, whilst models were adjusted for a range of clinical and sociodemographic features, several potentially important confounding factors that could influence the associations of metabolic traits with survival such as dietary intake, physical activity levels, time of blood draw, plasma sample storage time and medication use, could not be controlled for in the analysis because information on these factors were not available. The inclusion of BMI and SE variables in the hazard models may have captured some of the potential confounding by diet and physical activity, but it can be the case that someone who is very inactive and eats a poor diet may have a BMI in the normal range.

A fourth issue, which potentially affects the generalisability of my findings, relates to the fact that that those people with complete data (who were included in the analyses) were different at baseline compared to those with missing covariate data. These differences were apparent for comorbidity, age, HPV status and smoking status. I made an effort to examine whether this may have influenced my results by running the survival models in both the complete case dataset and the entire dataset (i.e. everyone with metabolomic data available) and plotting a linear fit of the overall correspondence in effect estimates. The results suggest that there is good correlation between HRs, as evidenced by an R_2 of 0.82 and a slope of 0.04. It is still possible however, that the sample used in my analysis is not representative of the OPC population as a whole.

A fifth limitation is that I was unable to validate my findings in an external OPC cohort. I am not aware of any other existing HNC cohorts with metabolomics data available at present, however, Nightingale Health has applied to UK Biobank to conduct metabolic profiling on all of its 500,000 participants using their high-throughput NMR platform [729]. A pilot study began in 2019 and is expected to take three years to complete. The number of HNC cases in this dataset will inevitably be smaller than the H&N5000 cohort, however.

Finally, whilst the number of biomarkers included in the NMR platform is relatively large compared to other existing platforms, it still only represents a fraction of the metabolites present in the blood. Metabolomics is a relatively new and emerging field of science and, at the present time, there is no single metabolomics technology or combination of technologies that can characterise the entire metabolome [730], largely because it incorporates such a diverse variety of endogenous and exogenous classes of compounds, with varying sizes, polarities and concentrations. It is likely that, with future advancements in high-throughput technologies enabling more biomarkers to be identified, many more mortality-associated biomarkers will be discovered, which may lead to improved risk prediction.

10.5.4. Conclusion

In conclusion, the findings of this analysis suggest that some metabolites within the metabolome may be related to overall mortality in people with OPC. A plasma metabolomic profile, characterised by high acetate levels in particular, may be associated with a poorer prognosis. These results should be validated externally in a large prospective cohort. Comparisons with the metabolomic profiles of healthy individuals should also be conducted to provide a better understanding of OPC aetiology.

Chapter 11: Discussion

11.1. Introduction

HNC develops due to an interplay of biological, lifestyle, and environmental factors. Similarly, many factors can affect an individuals' prognosis once they have developed the disease. These include features related to the cancer, (e.g. the type of cancer, where it originated in the body, and its stage), and features related to the individual themselves (e.g. their age, how healthy they were before they developed HNC and their lifestyle behaviours). This thesis explored the effects and possible prognostic value of a range of exposures on HNC survival, using a combination of self-report, epigenomic and metabolomic techniques. It was split into four separate but inter-related analyses. First, I examined whether smoking status and alcohol intake, factors which are known to associated with HNC risk, were observationally associated with mortality risk in a sub-sample of H&N5000 participants with oral, oropharyngeal, and laryngeal cancers. Second, I used robust and easy-to-measure DNAm-based predictors for these behaviours and two other complex traits, namely BMI and educational attainment, to examine the role of these exposures on survival in people with OPC. Third, I investigated whether DNAm-based predictors of aging, also termed "epigenetic clocks", provided prognostic information in the same sample of individuals with OPC. Finally, I looked at the potential relationships between circulating plasma metabolites and OPC cancer survival. To date, the prognostic significance of epigenetic and metabolomic factors have been largely overlooked in the HNC literature.

An in-depth discussion of was provided at the end of each results chapter. In this final chapter, I will summarise the key findings from my thesis, before evaluating the potential for better prognostic model development in HNC.

11.2. Summary of findings and implications

Table 46 summarises the main findings of this thesis.

Table 46: An overview of the research questions and findings of this thesis.

Research Question	n	Main findings	Chapter
What proportion of the variation in survival is explained by self-reported smoking and alcohol intake?	1,403	• Smoking explains an additional 7% of the variation in survival on top of age and gender (R_2=0.13 vs 0.06 for age and gender). • Alcohol intake explains an additional 2% of the variation in survival on top of age and gender (R_2=0.08 vs 0.06 for age and gender).	7
Is self-reported smoking status associated with all-cause mortality in people with HNC, after adjusting for clinical, biological and lifestyle factors?	1,403	• Compared to people who had never smoked, current smokers were twice as likely to die during follow-up (HR= 2.0 [95% CI:1.4, 3.0; *p* for trend <0.001]) • Former smokers, by comparison, were 60% more likely to die compared to never-smokers (HR=1.6 [95% CI:1.2, 2.3; *p*<0.001]. • There was no evidence of heterogeneity by treatment centre, which suggests that the results were not influenced by factors such as the preferred treatment approach at that hospital, level of care provided, or the general affluence/deprivation of the area.	7
Is self-reported alcohol associated with all-cause mortality in people with HNC, after adjusting for clinical, biological and lifestyle factors?	1,403	• There was no evidence that hazardous to harmful drinkers had an increased risk of all-cause mortality compared to non-drinkers in a model that adjusted for age, sex, stage, HPV status, comorbidity, marital status, annual household income and educational attainment (HR=1.0 [95% CI:0.8, 1.3; *p*=0.554]).	7

Table 46 continued.

Does cancer stage influence the association of self-reported smoking and drinking with all-cause mortality in HNC?	1,403 (low, n=581; high, n=822)	• Smoking status was associated with all-cause mortality in both high- and low-stage tumour groups (stages I and II, HR=2.8 [95% CI=1.3, 6.2; *p*=0.011]; stages III and IV, HR=1.9 [95% CI=1.2, 3.0; p=0.003]). • I found no evidence that alcohol drinking influenced all-cause mortality risk in either low- or high-stage cancer groups in fully adjusted models.	7
Does HPV status influence the association of self-reported smoking and drinking with all-cause mortality in people with OPC?	656 (HPV [-], n=176; HPV [+], n=480)	• There was weak evidence to suggest that hazardous to harmful drinkers with non-HPV associated cancers had worse survival compared to their non-drinking counterparts (HR=2.4 [95% CI:1.07, 5.59; *p* for trend =0.074]), but estimates were based on low sample numbers. • The analysis was underpowered to detect an effect of smoking status on all-cause mortality.	7
What proportion of phenotypic variance is explained by DNAm based predictors of smoking, alcohol drinking, BMI and educational attainment in people with OPC?	408	• DNAm predictors correlate with lifestyle factors that are associated with HNC risk and prognosis, but the proportion of the phenotypic variance explained differs. • Up to 48.65% of the variance in smoking and 16.0% of the variance in alcohol drinking can be explained by their respective predictors (AHRR and the score developed by Lui that includes 144 CpGs). • DNAm was found to explain 21.53% and 0.68% of the variation in BMI and educational attainment, respectively.	8

Table 46 continued.

What proportion of the variation in survival is explained by the DNAm-based predictors of smoking, alcohol drinking, BMI and educational attainment in people with OPC?	408	• On top of age and gender, DNAm-based smoking predictors increased the proportion of explained variation in survival by 5%-10% (R_2 =0.29-0.34 vs 0.24 for age and gender). • DNAm-based predictors of alcohol intake increased the proportion of explained variation in survival by 2%-3% (R_2 = 0.26-0.27 vs 0.24 for age and gender). • DNAm-based predictors of BMI and educational attainment increased the proportion of explained variation in survival by 2% (R_2=0.26 vs 0.24 for age and gender).	8
Are DNAm based predictors of smoking, alcohol intake, BMI and educational attainment associated with all-cause mortality in people with OPC?	408	• Four out of the five smoking-related DNAm scores considered (methylation at AHRR, both Joehanes scores and the score developed by Zhang) were associated with survival after controlling for age, gender, cell counts, batch effects, TNM, HPV status, comorbidity, income, marital status and self-reported smoking and drinking. • The highest effect estimate was obtained for AHRR (HR per unit increase in standardised DNAm score =1.92, 95% CI:1.06, 3.47).	8
What proportion of the variation in survival is explained by the DNAm-based predictors of EAA?	408	• The only age acceleration measure to increase the proportion of explained variation in survival (compared to the variation in survival explained by age and gender alone) was *AgeAccelGrim, (*R_2=0.36 vs 0.24 for age and gender).	9

Table 46 continued.

Are DNAm based predictors of epigenetic age acceleration associated with all-cause mortality in people with OPC, after controlling for clinical, biological and lifestyle factors?	408	• *AgeAccelGrim* (EAA based on GrimAge) and *IEAAHannum* (intrinsic EAA based on Hannum's clock) were associated with all-cause mortality in models that adjusted for gender, clinical, SE, and behavioural (smoking and alcohol intake). • HRs: 1.32 (95% CI: 1.08, 1.61; p=6.9 x10-03) and 1.39 (95% CI: 1.06, 1.83; p=0.017), for *AgeAccelGrim a*nd *IEAAHannum*, respectively.	9
Does the inclusion of a DNAm based measure of epigenetic age acceleration to a 'standard' clinical prognostic model improve 3yr mortality prediction?	408	• The addition of *AgeAccelGrim* to a standard clinical model that included TNM stage, HPV status, comorbidity and BMI, improved mortality prediction at 3 years, though the improvement was modest. • Clinical AUC: 0.77, clinical + AgeAccelGrim AUC: 0.80; *p*-value for difference=0.069.	9
Are circulating serum metabolic traits associated with all-cause mortality in people with OPC?	703	• Acetate, creatinine, omega-3 and histidine were associated with all-cause mortality in my primary analysis. • Each SD increase in acetate was associated with a 28% increased risk of death in the fully adjusted model. • Increases in circulating creatine, omega-3 and histidine were associated with improved survival (28%, 24% and 19% lower all-cause mortality risk per SD increase in metabolite), though *p*-values did not reach my threshold for multiple testing.	10

11.3. Strengths and limitations of this thesis

11.3.1. Strengths

11.3.1.1. Prospective study design

Unlike many other studies which have used historical cohorts or cancer registry data to examine the effects of smoking and alcohol intake on survival in HNC (see chapter 3), in this thesis I used data that was collected prospectively. Typically, prospective studies are more accurate with regards to the information collected about exposures, endpoints, and confounders and there is less missing data [731] [732]. This is because there is no need for any recollection of the information (i.e. less recall bias) and you can potentially go back to respondents if any information is missing, unlike in retrospective studies where reliable data on exposures or confounders may be unavailable or incomplete. In H&N5000, the study team went back to the recruitment site if there was information missing from a participant's data capture form. This reduced the amount of missing clinical data. Given that exposures were assessed at baseline, before my outcome of interest (death) had occurred, this also allowed me to calculate estimates of absolute risk (i.e. HRs). Another related advantage of the prospective cohort design is that individuals do not base their decision to take part in the study on their future outcome, since this is unknown. This reduces the potential for survivorship bias, which is a where people who do well are disproportionately evaluated. Focusing on survivors can result in a false, or disproportionate, estimate of probability or effect. Finally, cohort studies generally allow multiple exposures (and possibly outcomes) to studied simultaneously [732], making them particularly well-suited to prognostic research.

11.3.1.2. Availability of baseline clinical, biological and lifestyle data.

The availability of baseline data in H&N5000 is diverse and wide-ranging. As a result, I was able to look at the relationships of several exposures, including both biological exposures (DNAm levels) and modifiable behaviours (smoking and alcohol exposure), on prognosis. Moreover, I was able to investigate the impact of confounding on these relationships due to the rich amount of data available on socio-demographic and clinical variables. As illustrated in chapter 3, the majority of studies looking at the role of smoking and alcohol consumption in HNC survival did not adjust for, or were unable to adjust for, HPV status, comorbidity or

BMI (often because they were conducted retrospectively), factors which are known or purported to impact on mortality risk. In this thesis, I found that controlling for these variables in survival models led to an attenuation of the effect sizes. This may explain why I found no evidence that pre-treatment levels of alcohol consumption influenced survival in my analysis whilst Duffy and others [441] [442] [447] found that active or heavy-drinkers experienced worse survival.

11.3.1.3. Data linkage

A major advantage of the H&N5000 study design for survival analysis is that all consenting participants were linked to NHS digital at the start of the study, meaning that even if the study sites were unable to collect data from participants medical notes, the study team would be notified of any subsequent deaths among cohort members. This minimises loss to follow-up. Loss to follow-up is a concern in any clinical study as it reduces the power of a statistical analyses and has the potential to introduce bias if the individuals who are loss to follow-up differ in any way from those that remain in the study (a problem termed attrition bias).

At the time of writing this thesis, 106 people were known to have withdrawn from the H&N5000, but only seven of these had withdrawn from data linkage. Since the exposures in my analyses were measured at baseline and my dataset only included people with baseline data, any withdrawals from follow-up questionnaires would not have influenced my data. The seven people who withdrew from linkage, for whom my event (death) had not occurred, would have been censored in my dataset, but this is unlikely to have influenced my overall findings, given my overall sample size.

11.3.1.4. Sample size relative to other HNC studies

H&N5000 is one of the largest prospective clinical cohorts of people with HNC in the world, with a wide range of information available to researchers (biological, clinical and patient-reported). Of note, very few HNC cohorts, if any, have accompanying metabolomic data. As such, the metabolic perturbations occurring during the pathogenesis of HNC and their correlation with clinical outcomes is generally not well studied. In my scoping review (Chapter 3; table x), the largest metabolomics study I found included 159 OSCC samples. This included tumour tissues, neighbouring margins, and bed tissues. The largest study in blood included 140 HNC serum samples. On this basis, my analysis is the largest of its kind to date, with a final sample size of 703. In addition, most of the studies I identified compared

the metabolomes of people with and without HNC, with the view to identifying diagnostic or screening biomarkers, rather than investigating the prognostic role of metabolites. My analysis therefore addressed a gap in the existing literature.

11.3.2. Weaknesses

11.3.2.1. Selection into the H&N5000 study

It was estimated that when all study centres were open, H&N5000 captured a third of all incident cases in the UK. It is rare for selection bias to occur in cohort studies as bias only arises when a factor causing bias (e.g. inclusion) is related to *both* the exposure and the outcome. As outcome is in the future in a prospective cohort like HN5000 and participants did not base their decision to take part in the study upon the future outcome, selection forces may not actually bias. However, it is conceivable that this sample is not representative of the HNC population as a whole. The decision to participate may correlate with social, educational and health circumstances and these factors may themselves correlate with risk factors for mortality. Generally speaking, non-participation and tend to occur more commonly among less affluent, less healthy people and this can result in cohort studies being made up of relatively healthy, affluent sub-populations [733].

Of note, the majority of participants in H&N5000 are white. People of other ethnicities may have been less likely to be recruited into the study due to language barriers, or alternatively, white people may be more likely to participate in research. Factors such as smoking status and BMI may have a greater or lesser influence on mortality risk in different ethnicities, due to genetic variation.

Taken together, these issues may affect generalisability. Further studies, using larger numbers of subjects of all ethnicities are required in order to establish whether my findings can be generalised to the HNC population as a whole.

11.3.2.2. Selection into the analytic sample

Whilst selection bias is unlikely to be problematic in H&N5000 overall, it may be an issue in my thesis because individuals were only included in the analyses if they had baseline blood samples and questionnaire data available. If the decision to participate in baseline data collection is correlated with risk factors for the outcome under study (mortality), then possible

bias due to baseline selective participation cannot be ruled out. This issue may be further compounded in complete case analyses.

11.3.2.3. Missing covariate data

Some of the covariates considered in this thesis had missing data, either because baseline questionnaires were not fully completed by some of the respondents (intentionally or by mistake) or because necessary information was not collected at the start of the study (i.e. height and weight). This presented various problems. For instance, most statistical packages will eliminate cases when they encounter missing data for any of the variable included in the analysis, which in turn reduces statistical power. So, even if each individual variable only has a small amount of missing data, when examined in combination with the other variables included in the model, this can result in a drastic reduction in the number of individuals available for analysis. Another potential problem is that missing data can introduce bias, giving rise to misleading results. For instance, in my analysis, annual household income had some of the highest amounts of missing data. If people who were very affluent chose not to disclose their income, they would be under-represented in the sample and my results could be different, perhaps because people who are more well-off have better survival, irrespective of lifestyle exposures.

In my thesis, I attempted to minimise the effects of missing data by performing MI, which involves creating several simulated complete versions of the data set, analyses each of these new data sets separately, and pooling the results [672]. The advantage of using this approach is that it makes use of all the available data and thus preserves sample size. As discussed in my results chapters however, MI will only produce unbiased estimates in situations where the data is MCAR or MAR; if the reasons for missing data are related to the variables under investigation, this indicates sampling bias and MI would not appropriate. Given that there is no formal test for establishing the mechanism of missingness and that sampling bias could not be ruled out in my analysis, I also conducted complete-case analyses and compared the results. Overall, I found that the results of my imputed and complete-case analyses were largely comparable.

11.3.2.4. Power and sample size

Statistical power is an important consideration when designing prospective studies [734]. Power calculations are used to determine how many study participants are needed in order to avoid a type II ("false negative") error [735]. The "power" of the study is equal to 1- β, where β is the proportion of results that were incorrectly reported as being negative. Most studies accept a power of 80%, meaning that they are willing to miss a real difference or effect 20% of the time.

The H&N5000 study was calculated as having 80% power to detect a difference in survival of around four percentage points for an intra-class correlation coefficient (ICC) of 0.005 [574], based on a sample size of 4,000 people (allowed for exclusions of rarer cancer types, withdrawals from the study, incomplete data and loss to follow-up from the target total of 5,000). This power calculation was based on the aim of the study, however, which was to evaluate the impact of centralisation of care for people with HNC; therefore, its relevance to my analyses is limited.

In my thesis, which only included a sub-sample of H&N5000 participants, I did not determine the power of the analyses performed but instead used all of the data that was available. This is limiting because it is difficult to know whether I can trust a negative result, that is whether the test had sufficient statistical power to detect an effect if it exists. I could have performed a post-hoc power calculation, using the observed effect size and one of several free on-line calculators (e.g. G* Power, OpenEpi, EpiTools) [736]. However, the use of retrospective power analysis has been heavily criticised in the literature [734 737-740] and on online statistical forums, though editors often ask authors to include them. There are two main arguments to support this. Firstly, post hoc power calculations assume that the observed effect is similar to the true effect, but in reality, the true values are typically unknown. Secondly, as explained by Hoenig & Heisely [738], observed (post-hoc) power and *p*-values are inextricably linked because the observed significance level of a test, i.e. the *p*-value, also determines the observed power. As such, non-significant *p*-values will always correspond to low observed powers and observed power will add nothing to the interpretation of results. Overall, it seems that observed power is a tautology because the question you are asking is "What was the probability of finding a statistically significant result, assuming that the actual effect is the same as the observed effect?" but the results of the study already exist, so there is no real likelihood of producing such a statistically significant result [739 741]. Plate *et al* liken this to asking what the probability of winning the lottery is, when you already know that you have not won [741]. Goodman and others suggest that CIs are more useful for evaluating "non-

significant" results than host-hoc power calculations [739], since the width of the CI gives an indication of the likelihood of the true (population) effect size being equal to zero.

11.3.2.5. Lack of cancer-specific mortality data

The end-point or outcome of interest in my thesis was death by any cause. I would have liked to have also examined the effect of my exposures on HNC-specific mortality, but this data was not available at the time of my analyses. Ideally, cause of death would be determined by a panel of experts (e.g. clinicians, oncologists, pathologists) with access to the individuals' medical history, but this requires a considerable investment in time and resources. A project is currently underway to determine cause of death in H&N500, so future studies may be able to examine the role of biological and lifestyle traits on HNC-specific mortality.

Both outcomes, i.e. all-cause mortality and disease-specific mortality, have their advantages and disadvantages. Cancer-specific mortality is sometimes criticised because of possible biases in the determination of cause of death [742]. For example, based on evaluation by expert panels and/or autopsy reports, 15-35% of death certificates are estimated to misclassify cause of death [742]. Overall mortality is, by contrast, an objective measure. However, where the study population has high mortality rates from co-occurring conditions, as is the case with HNC, differences in survival between groups due to the exposure of interest may be overshadowed by deaths due competing causes e.g. cardiovascular disease, chronic obstructive pulmonary disease and second primary cancers in the case of HNC [743]. Given that a considerable proportion of individuals may die from causes other than their HNC, larger sample sizes are however needed to detect an effect of the exposure on cancer-specific mortality.

Ultimately, a person with disease may be less concerned about what they will die of and more about their chance of survival, in which case all-cause mortality may be a better study outcome.

11.3.2.6. Sample collection and handling

The initial overall aim of the H&N5000 study was to evaluate the outcome of centralisation in HNC. As such, whilst one of the objectives was to create a resource for translational and applied research in HNC, the study was not designed specifically to conduct metabolomic

analyses. The opportunity to do so was born out of the success of the study and the research team. The upshot of this was that sample collection and handling procedures were not well-suited for metabolomic studies. Most significantly, the samples were not immediately frozen following blood draw and they were posted to the study laboratory at ambient temperature. As a result, I was not able to use all of the NMR metabolite data, including glycolysis-related measures, because of sample degradation. I would have been interested to investigate the role of circulating biomarkers of glycolysis, given that cancer cells, in general, show an increased dependence on aerobic glycolysis (termed the Warburg effect) [744].

11.3.2.7. Limited epigenetic and metabolomic data

Epigenetic and metabolomic data are currently unavailable for the entire H&N5000 cohort, limiting the sample size in my analyses and my ability to stratify by cancer site and HPV status. Samples were selected for epigenetic profiling based on a clinical ICD code of OPC and the availability of baseline questionnaire data, data capture and biological samples. This meant that, of the 1,896 individuals who received an initial diagnosis of OPC, epigenetic data was only available for 445 of them. Five samples had to then be excluded because they failed QC and a further 32 were removed from my analysis because the cancers were recoded based on further pathological information.

I obtained metabolomic data for all of the OPC cases with adequate blood available. This gave me an initial sample size of 1,595. However, 112 cases were removed as a result of pathological re-coding. Whilst I was unable to look at the effect of the plasma metabolome on survival in other HNC sites, this sample size is still large relative to other metabolomic studies in the HNC literature, as highlighted above.

11.3.2.8. External validity

The uptake of biological biomarkers in clinical practice requires vigorous validation in external settings. It was not possible to validate the findings of my PhD in an external population because I am unaware of any other existing prospective HNC cohorts with epigenetic, metabolomic and mortality data currently available. The extent to which my findings can be applied to other settings and other populations cannot therefore be determined at present, though I attempted to adjust some of my results for over-fitting. To advance reproducibility in different populations, further longitudinal studies are required.

11.4. Future work

There is a lot of scope to extend my thesis and develop a formal HNC prognostic model. Below I outline some of the possible future projects that would be interesting and potentially informative.

11.4.1. Mendelian randomization study of metabolite profiles

My metabolomics analysis was hypothesis generating in design. It uncovered several metabolites that might be able to predict mortality risk in HNC, but whether or not these biomarkers are causally related to HNC survival was not established.

Mendelian randomization (MR) provides an opportunity to test for causal effects between the metabolites I identified in my analysis and risk of death. MR is a type an instrumental variable analysis which is being increasingly used in epidemiological research [745]. It uses genetic variants robustly associated with your trait of interest (here, my metabolite traits) as 'proxies' for that exposure. The underlying principle or assumptions of MR are that the allocation of genetic variants (alleles) from parent to offspring are random and unrelated to factors other than that trait, that is, they are not affected by confounding [746]. In this way, MR is often described as 'nature's randomized controlled experiment' because in MR people are allocated to different exposure levels ("treatment groups") based on their genetic make-up, which is randomised at conception [747]. This randomisation event ensures that confounders are distributed approximately equally across exposure groups, irrespective of whether or not they are measured or whether they have been measured accurately.

In the future, I would look to employ a two-sample MR approach to estimate the causal effect of metabolite concentrations on HNC survival [747]. This means that the effects of the instrumental variable-risk factor association (i.e. the single-nucleotide polymorphisms (SNPs) - exposure association) and the instrumental variable-outcome association (i.e. the SNP- outcome association) would not have to be obtained from the same set of individuals. Instead, I could make use of publicly available GWAS summary data. This approach has several advantages over one-sample MR, as discussed in Lawlor (2016), including increased power and the fact that in two-sample MR, weak instrument bias is towards the null [747].

I would first have to identify appropriate genetic instruments for acetate, creatinine, omega-3 and histidine that can be used within an MR framework, which I could do using the on-line platform MR-base [748] [749]. The literature shows that there are strong associations between SNPs and metabolite levels [597] [750], and therefore these SNPs could be used in an MR framework, providing they are in linkage disequilibrium (i.e. they are independent). For instance, Kettunen et al (2016) observed that the proportion of variance explained for metabolite- SNP associations in their analysis ranged from 0.2% for acetoacetate to 12.5% for glycine, with a median of 5% [597].

Approaches such as MR required very large sample sizes and at present, there are limited HNC datasets with genome-wide genetic available for use. Collaborative efforts, including those of the International Head and Neck Cancer Epidemiology Consortium (INHANCE), a collaboration of over 40 studies with approximately 40,000 cases and controls, are ongoing. That said, many of the studies contributing to INHANCE are relatively old. Further genotyping in newer cohorts such as HN5000 are hopefully in the pipeline. Indeed, funding to complete genotyping of over 2,500 HNC cases (oral, oropharyngeal and laryngeal cancers) from HN5000 has been secured by Dr Tom Dudding at the University of Bristol, who is also working closely with collaborators at IARC to obtain funding to genotype a further 5,351 HNC cases and 4,351 controls.

Of note, MR for survival analysis can be complicated by issues of collider bias, that is bias in a measure of association between two variables due to conditioning on a common effect of exposure and outcome [751] [752]. This is less of a problem if you are interested in prediction, however, rather than causal association.

11.4.2. Saliva and tissue metabolites and survival in HNC

I identified several candidate biomarkers for prognosis in plasma. In addition to blood samples, the H&N5000 bioresource also holds 4,899 pre-treatment saliva samples (oral cavity, n=1,199; oropharynx, n=1,752; larynx, n=986) and 2,518 surgically resected tissue samples (oral cavity, n=710; oropharynx, n=839; larynx, n=486), offering the potential for further metabolomic analyses, though there may be similar limitations around storage and handling of saliva as there were for blood samples. The analysis of the salivary metabolomic profile, in particular, presents an attractive alternative for prognostic research given its close proximity to tumour site [753]. In addition, saliva is readily available and it can be collected easily and non-invasively, compared to blood collection which requires a trained

phlebotomist [754]. Saliva is a very complex biofluid. It contains a cocktail of secretions from the major and minor salivary glands and gingival crevicular fluid, as well as serum and blood derivatives from oral wounds, desquamated epithelial cells, bronchial and nasal secretions, bacteria and bacterial products, viruses and fungi [755 756]. As such, it is considered an important source of biological information. Previous studies have proposed using saliva metabolites to differentiate between pre-cancerous and malignant HNC lesions [271], but so far, studies have yet to determine if they can provide prognostic value.

11.4.3. HNC prognostic model development

Variable selection is a fundamental issue when developing any prognostic model. In this thesis, I used evidence from the existing literature and survival analysis techniques to identify a list of candidate predictor variables. The next obvious step in my investigation would be to develop a parsimonious prediction model that includes as few predictor variables as possible, i.e. a model that accomplishes the best prediction with as few parameters as possible. An over-complicated model including unreliable variables would not be adopted in clinical practice.

There are a number of automated model selection methods available, including backward-elimination, forward selection, stepwise forward/backward algorithms, information criterion and penalized likelihood methods such as LASSO and elastic net regression [757]. The latter methods of which are popular in high-dimensional model selection (e.g. the selection of metabolite variables [758 759]).

11.4.4 Application and impact of a new HNC prognostic model in clinical practice

Before embarking on developing an entirely new prognostic model, it is important to consider whether it is likely to be useful in clinical practice. As pointed out in Chapter 2, prognostic models are relatively abundant in the HNC literature; yet whilst multiple prediction models are being developed to predict health outcomes in this population, the exploitation of lifestyle and molecular data for the improvement of prognosis and treatment selection has not yet become routine, with the exception of p16 expression. There are a number of reasons why this may be the case. Firstly, before any model can be adopted in clinical practice, it is vital that it is shown to perform well (i.e. provide accurate predictions) outside the specific context of the sample in which it was developed, yet only a small number of studies report external validation. Secondly, the addition of new markers may yield only

moderate benefit given that standard models generally include the most important or influential predictors. Finally, models are only practical if clinicians can easily obtain the data required to make their prediction.[467]

With these factors in mind, I propose that, rather than attempting to develop a new HNC model from scratch in H&N5000, it would be better to build upon existing HNC models to see whether they can be improved or recalibrated. Indeed, the Prognosis Research Strategy (PROGRESS) series recommends that "rather than developing a steady stream of new prognostic models, researchers should shift to validation, updating, and impact studies of existing models" [464].

The performance of existing models may diminish over time if, for example, HNC cases are getting identified earlier or if treatment changes. Additionally, new markers may become available. In this current era of biomarkers and "omics," assessment of the extent to which novel markers add value to existing models is increasingly important.

Another possible motivation for updating a model may be that existing predictors are prone to substantial measurement error or reporting bias (e.g. smoking status). I would look to consider whether existing prognostic models could be improved by adding the novel biomarkers identified in this thesis, namely DNAm based measures of smoking, alcohol intake and biological aging.

As stated above, a model is only likely to be adopted by clinicians if the data needed to make predictions is readily available. A limitation of using the epigenetic predictors from this thesis is that DNAm profiling is not routinely collected at the time of a HNC diagnosis. Therefore, the addition of peripheral, blood-based DNAm scores to standard clinical prognostic models would have to add substantial prognostic information in order to advocate changing clinical practice. Whilst more affordable methods for DNAm analysis are becoming available, cost remains a significant factor for the Health Service. Moreover, blood collection and handling is time-consuming and adds an additional burden to both the clinician and the individual with HNC. If however, an easy to measure DNAm-based predictor of lifestyle exposure did improve prognostic model performance sufficiently, this could have the potential to play an important role in pathways towards improved health, including clinical decision-making and the development, evaluation, and targeting of interventions in people with HNC.

11.5. Concluding remarks

In summary, this thesis demonstrates the potential to improve upon existing HNC prognostic models by integrating epigenetic and metabolomic technologies. In particular, I have shown that DNAm-based biomarkers for smoking, and to a lesser extent epigenetic aging, may augment prognostication in this population. The ability of biological predictors to inform on (potentially modifiable) exposures could lead to an improved understanding of disease variance and individual risk and help patient stratification. Existing DNAm predictors for alcohol, BMI and educational attainment do not appear to provide added prognostic information, beyond established risk factors such as cancer stage and HPV status, but the phenotypic variance explained by these biomarkers is limited. Future EWAS studies may identify additional CpG sites whose methylation levels are related to these exposures, which would in turn benefit subsequent epidemiological studies of disease risk and prognosis.

The development of metabolomic biomarkers for HNC prognosis is at present hampered by the lack of standardised analysis approaches and operating procedures for use in sample collection and handling. However, the feasibility of metabolomics for biomarker discovery in this area is supported by the knowledge that malignant transformation causes disruption of biochemical pathways, fuelling disease progression. Further, well-designed studies in different biological samples along with internal and external validation of study findings are needed to elucidate the prognostic role of metabolites in HNC.

Appendix A

A1: Baseline descriptives of people included in the imputed analysis (n=3,890).

	Oral cavity (n=1150)		**Oropharynx (n=1770)**		**Larynx (n=970)**		
Characteristic	*N*	*Frequency*	*N*	*Frequency*	*N*	*Frequency*	*p-value**
Gender							
Male	707	61.50%	1403	79.30%	832	85.80%	
Female	443	38.50%	367	20.70%	138	14.20%	<0.001
TNM stage							
I	390	34.00%	74	4.20%	400	41.20%	
II	268	23.40%	174	9.90%	246	25.40%	
III	92	8.00%	254	14.40%	161	16.60%	
IV	397	34.60%	1261	71.50%	163	16.80%	<0.001
HPV status							
Negative	943	96.60%	455	30.60%	798	97.40%	
Positive	33	3.40%	1032	69.40%	21	2.60%	<0.001
Comorbidity							
None	431	38.20%	834	47.90%	348	36.70%	
Mild	401	35.60%	595	34.20%	347	36.60%	
Moderate/severe	295	26.20%	312	17.90%	253	26.70%	<0.001
Education							
School education	382	46.60%	562	44.30%	389	58.10%	
College	281	34.30%	465	36.70%	198	29.60%	
Degree	156	19.00%	241	19.00%	83	12.40%	<0.001
Annual household income							
<£18,000	367	51.00%	449	38.10%	360	59.20%	
£18000-£34,999	202	28.10%	358	30.40%	159	26.20%	
>£35,000	151	21.00%	372	31.60%	89	14.60%	<0.001
IMD							
Low Deprivation	406	39.30%	638	39.40%	401	46.00%	
Moderate Deprivation	216	20.90%	358	22.10%	192	22.00%	
High Deprivation	410	39.70%	623	38.50%	279	32.00%	<0.003
Relationship status							
single (never married)	129	14.90%	147	11.20%	77	10.80%	
currently in relationship	542	62.50%	935	71.30%	484	67.60%	
No longer with spouse	196	22.60%	229	17.50%	155	21.60%	<0.001
Smoking status							
Never	208	25.00%	348	27.50%	65	9.40%	
Former	430	51.70%	696	55.00%	491	71.00%	
Current	194	23.30%	221	17.50%	136	19.70%	<0.001
Alcohol intake							
non-drinker	230	27.30%	317	24.50%	185	26.60%	
moderate drinker	176	20.90%	285	22.10%	151	21.70%	
hazardous-harmful drinker	435	51.70%	690	53.40%	360	51.70%	0.660
	N	*Mean (SD)*	*N*	*Mean (SD)*	*N*	*Mean (SD)*	
Age (years)	1139	62.2 (12.3)	1759	58.9 (9.0)	969	65.1 (10.4)	<0.001
Body mass index (kg/m2)	670	26.0 (5.2)	1032	26.9 (5.2)	589	26.3 (5.2)	0.002

A2: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by tumour site (n=3,890).

	All sites (n=3,890)				**Oral cavity (n=1150)**				**Oropharyngeal (n=1,170)**				**Larynx (n=970)**			
	HR	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**
Model 1																
Smoking																
Former	1.66	1.34	2.06		1.54	1.12	2.12		1.95	1.40	2.73		2.00	1.02	3.89	
Current	3.18	2.55	3.98	<0.001	2.08	1.46	2.99	<0.001	4.31	2.97	6.26	<0.001	4.36	2.13	8.89	<0.001
Alcohol amount																
Moderate	0.89	0.73	1.09		0.86	0.62	1.21		1.01	0.72	1.41		0.78	0.51	1.19	
Hazardous/harmful	1.21	1.02	1.44	0.012	1.09	0.83	1.44	0.452	1.38	1.02	1.85	0.021	1.06	0.74	1.53	0.608
Model 2																
Smoking																
Former	1.56	1.25	1.96		1.63	1.17	2.27		1.66	1.17	2.35		1.76	0.88	3.49	
Current	2.20	1.74	2.80	<0.001	1.84	1.25	2.73	0.002	2.37	1.53	3.66	<0.001	3.11	1.51	6.39	<0.001
Alcohol amount																
Moderate	0.89	0.72	1.09		0.83	0.59	1.16		1.05	0.74	1.49		0.73	0.47	1.14	
Hazardous/harmful	1.09	0.91	1.30	0.207	1.00	0.75	1.34	0.868	1.18	0.88	1.58	0.245	1.03	0.70	1.51	0.689
Model 3																
Smoking																
Former	1.49	1.19	1.88		1.61	1.15	2.25		1.55	1.08	2.23		1.73	0.86	3.47	
Current	1.98	1.54	2.55	<0.001	1.71	1.14	2.56	0.006	1.98	1.25	3.13	0.003	3.16	1.53	6.56	<0.001
Alcohol amount																
Moderate	0.92	0.75	1.14		0.85	0.60	1.20		1.16	0.81	1.67		0.75	0.48	1.18	
Hazardous/harmful	1.13	0.94	1.35	0.102	1.00	0.75	1.33	0.861	1.30	0.95	1.76	0.085	1.09	0.73	1.64	0.591
Model 4																
Smoking																
Former	1.48	1.17	1.87		1.63	1.16	2.31		1.53	1.06	2.21		1.71	0.85	3.44	
Current	1.94	1.51	2.50	<0.001	1.74	1.14	2.64	0.006	1.94	1.23	3.08	0.004	3.09	1.50	6.40	<0.001
Alcohol amount																
Moderate	0.91	0.74	1.13		0.81	0.58	1.15		1.17	0.82	1.67		0.75	0.48	1.18	
Hazardous/harmful	1.06	0.89	1.28	0.321	0.90	0.67	1.22	0.697	1.26	0.93	1.71	0.132	1.04	0.69	1.58	0.776

*Model 1: adjusted for age and gender; Model 2: additionally adjusted for clinical features (TNM stage, comorbidity, BMI, HPV); Model 3: additionally adjusted for social features (education, annual household income, IMD and marital status); Model 4: additionally includes smoking or drinking. * test for linear trend. Values with p<0.05 are shown in bold. Abbreviations:* ***Haz****, hazardous;* ***HR****, hazard ratio;* ***Lower CI****, lower confidence interval;* ***Upper CI****, upper confidence interval.*

A3: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by tumour stage.

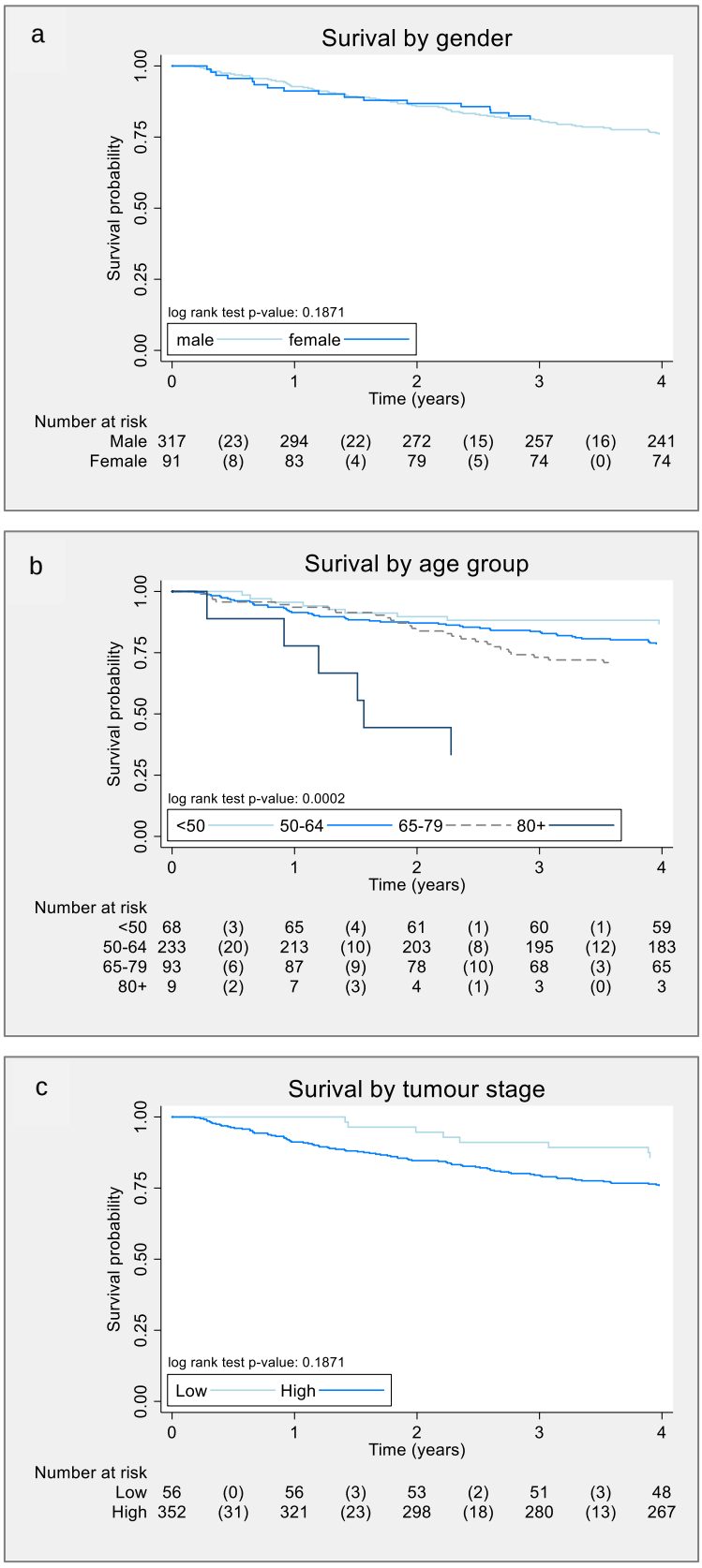
	Low stage				High stage			
	HR	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**
Model 1								
Smoking								
Former	2.01	1.26	3.20		1.69	1.33	2.15	
Current	3.68	2.34	5.80	<0.001	3.37	2.62	4.34	<0.001
Alcohol amount								
Moderate	0.81	0.56	1.18		0.90	0.70	1.14	
Hazardous/harmful	1.11	0.81	1.52	0.413	1.25	1.01	1.53	0.015
Model 2								
Smoking								
Former	1.94	1.23	3.07		1.40	1.09	1.79	
Current	2.96	1.85	4.72	<0.001	1.95	1.48	2.57	<0.001
Alcohol amount								
Moderate	0.82	0.58	1.17		0.91	0.71	1.18	
Hazardous/harmful	1.07	0.78	1.46	0.554	1.13	0.91	1.40	0.116
Model 3								
Smoking								
Former	1.83	1.15	2.91		1.33	1.04	1.72	
Current	2.64	1.63	4.28	<0.001	1.76	1.31	2.36	<0.001
Alcohol amount								
Moderate	0.87	0.61	1.25		0.93	0.72	1.21	
Hazardous/harmful	1.12	0.81	1.55	0.391	1.15	0.92	1.42	0.115
Model 4								
Smoking								
Former	1.82	1.14	2.90		1.32	1.02	1.70	
Current	2.60	1.60	4.21	<0.001	1.71	1.28	2.30	<0.001
Alcohol amount								
Moderate	0.85	0.59	1.22		0.93	0.72	1.21	
Hazardous/harmful	1.03	0.75	1.41	0.718	1.10	0.89	1.37	0.249

A4: Mortality hazard ratios (HR) according to baseline smoking and drinking status stratified by HPV status.

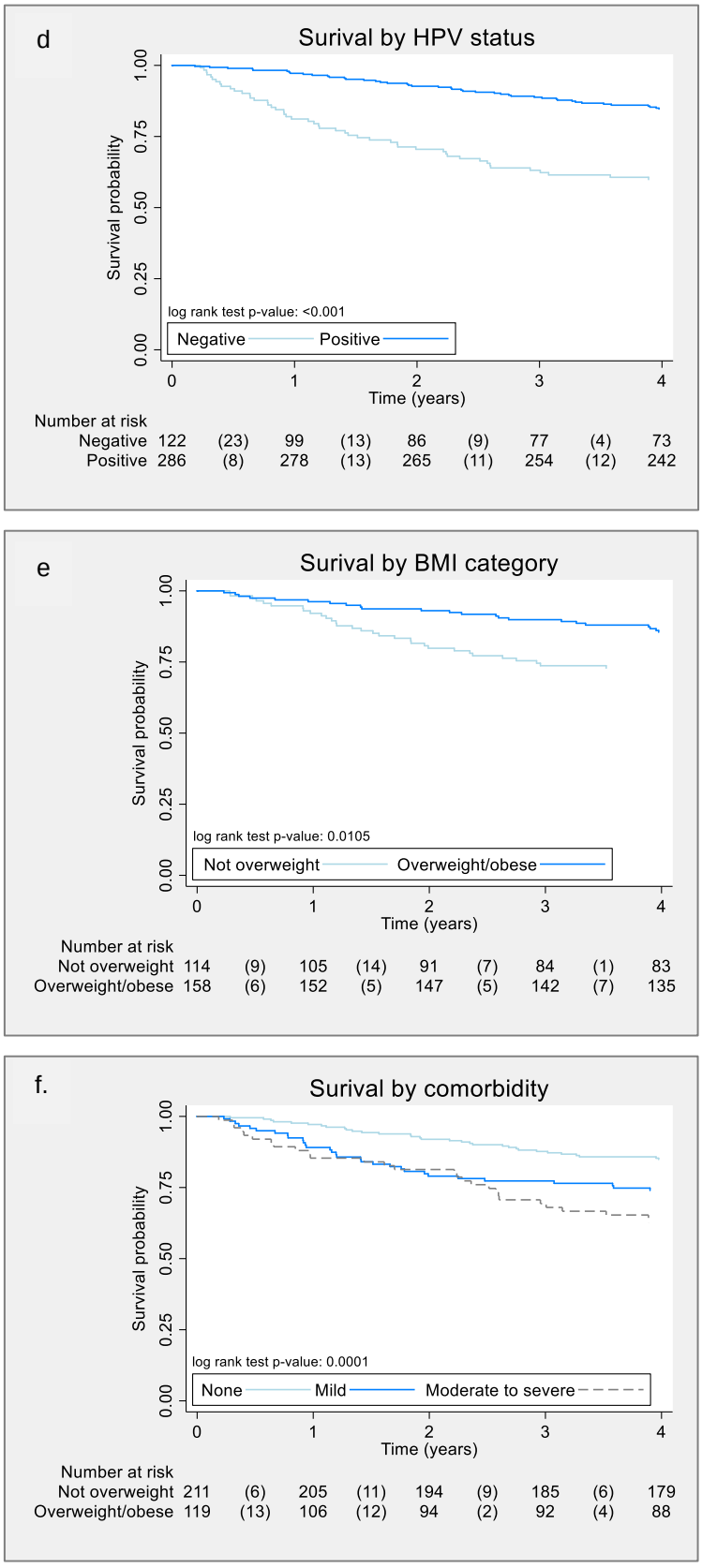
	HPV negative				HPV positive			
	HR	*Lower CI*	*Upper CI*	*p-value**	*HR*	*Lower CI*	*Upper CI*	*p-value**
Model 1								
Smoking								
Former	2.81	1.09	7.24		1.46	0.99	2.15	
Current	4.10	1.59	10.56	0.001	2.44	1.37	4.34	0.003
Alcohol amount								
Moderate	1.03	0.59	1.79		0.95	0.61	1.48	
Hazardous/harmful	1.35	0.87	2.09	0.136	1.03	0.70	1.53	0.814
Model 2								
Smoking								
Former	2.54	1.00	6.46		1.46	0.98	2.15	
Current	3.25	1.26	8.40	0.011	2.28	1.28	4.08	0.005
Alcohol amount								
Moderate	1.05	0.61	1.82		1.02	0.65	1.60	
Hazardous/harmful	1.26	0.82	1.94	0.242	1.04	0.70	1.56	0.836
Model 3								
Smoking								
Former	2.31	0.91	5.88		1.42	0.95	2.12	
Current	2.65	1.03	6.84	0.043	1.97	1.08	3.60	0.018
Alcohol amount								
Moderate	1.17	0.65	2.09		1.12	0.71	1.78	
Hazardous/harmful	1.37	0.86	2.18	0.165	1.16	0.77	1.76	0.416
Model 4								
Smoking								
Former	2.20	0.85	5.66		1.42	0.94	2.13	
Current	2.52	0.97	6.50	0.057	1.98	1.07	3.64	0.020
Alcohol amount								
Moderate	1.14	0.64	2.04		1.15	0.73	1.83	
Hazardous/harmful	1.30	0.82	2.05	0.220	1.15	0.76	1.74	0.504

Appendix B

B1: Kaplan-Meier plots of overall survival in univariate models (n=408).

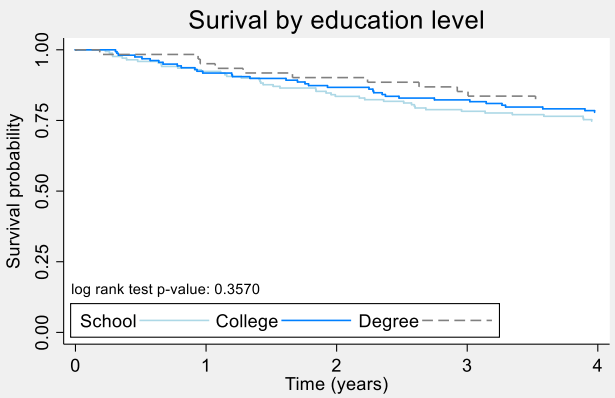


B1 continue.


d
Surival by HPV status
Survival probability
0.00
0.25
0.50
0.75
1.00
log rank test p-value: <0.001
Negative
Positive
0
1
2
3
4
Time (years)
Number at risk
Negative 122 (23) 99 (13) 86 (9) 77 (4) 73
Positive 286 (8) 278 (13) 265 (11) 254 (12) 242
e
Surival by BMI category
Survival probability
0.00
0.25
0.50
0.75
1.00
log rank test p-value: 0.0105
Not overweight
Overweight/obese
0
1
2
3
4
Time (years)
Number at risk
Not overweight 114 (9) 105 (14) 91 (7) 84 (1) 83
Overweight/obese 158 (6) 152 (5) 147 (5) 142 (7) 135
f.
Surival by comorbidity
Survival probability
0.00
0.25
0.50
0.75
1.00
log rank test p-value: 0.0001
None
Mild
Moderate to severe
0
1
2
3
4
Time (years)
Number at risk
Not overweight 211 (6) 205 (11) 194 (9) 185 (6) 179
Overweight/obese 119 (13) 106 (12) 94 (2) 92 (4) 88

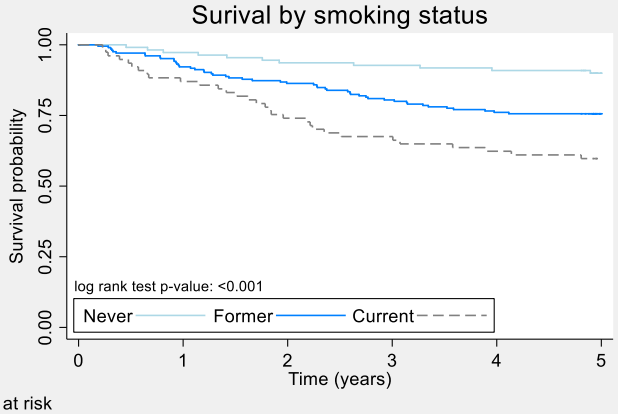
B1 continued.

g



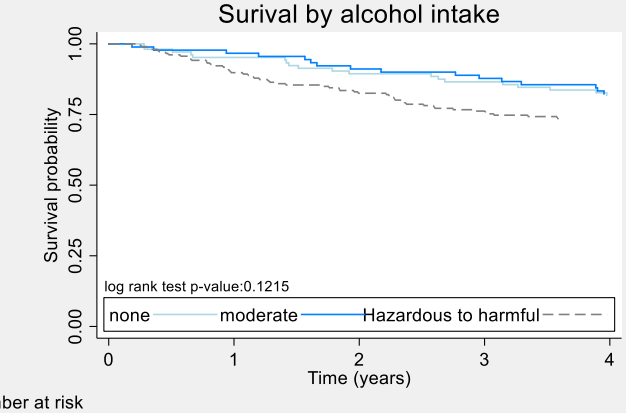
Number at risk									
School	170	(13)	157	(15)	142	(9)	133	(6)	127
College	158	(13)	145	(8)	137	(7)	130	(7)	123
Degree	61	(3)	58	(3)	55	(3)	52	(2)	50

h



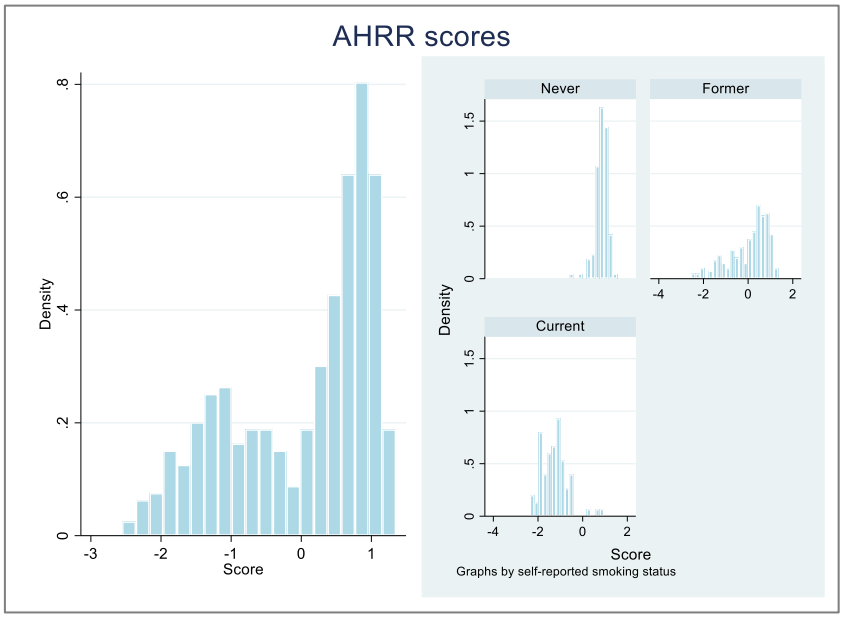
Number at risk											
Never	110	(3)	107	(4)	103	(1)	102	(2)	100	(1)	86
Former	205	(16)	189	(12)	177	(12)	165	(9)	156	(1)	130
Current	77	(10)	67	(10)	57	(5)	52	(4)	48	(2)	39

i.

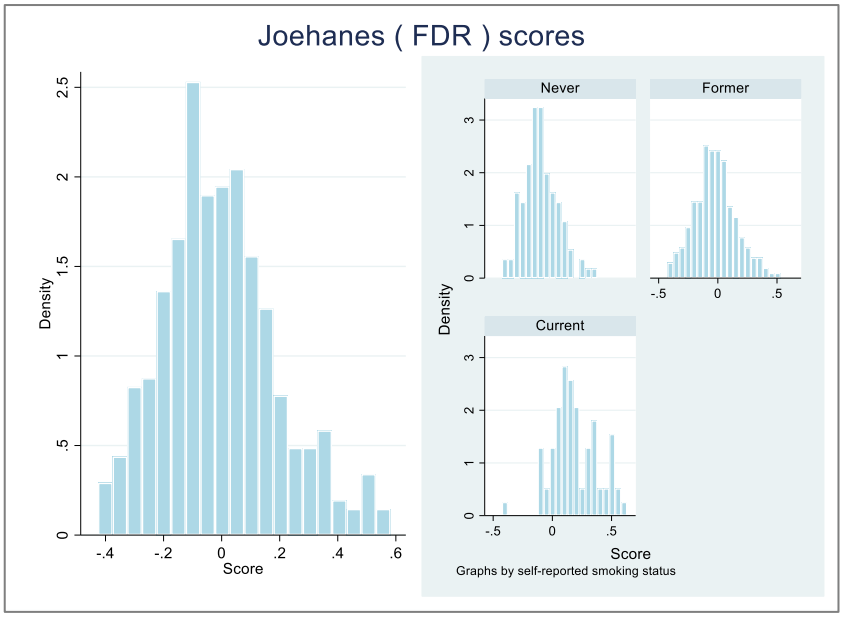


Number at risk									
Non	104	(5)	99	(6)	93	(3)	90	(5)	85
Moderate	90	(3)	87	(5)	82	(3)	79	(5)	74
Hazardous-harmful	206	(21)	185	(15)	170	(13)	157	(6)	151

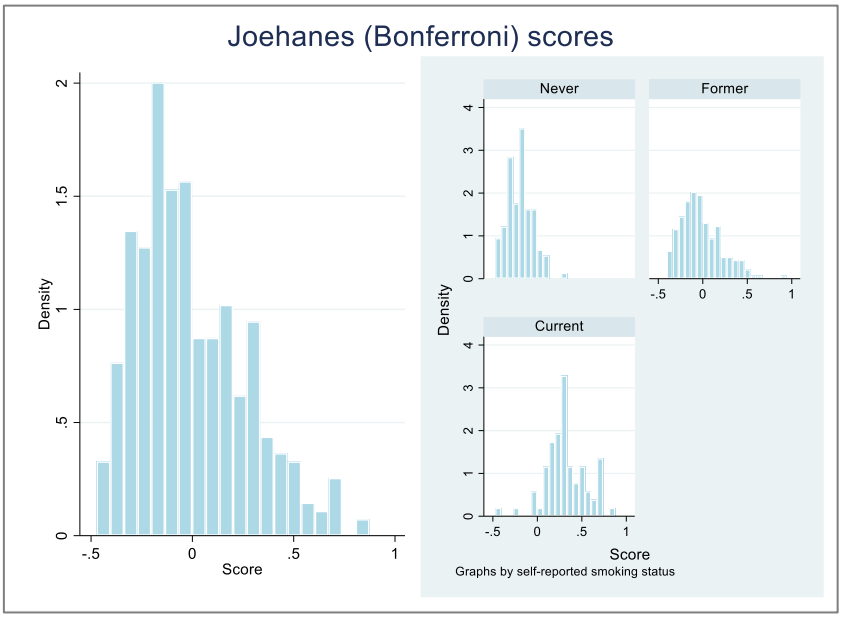
B2: Histograms showing the distribution of AHRR DNAm scores, overall and by self-reported smoking status (n=408).



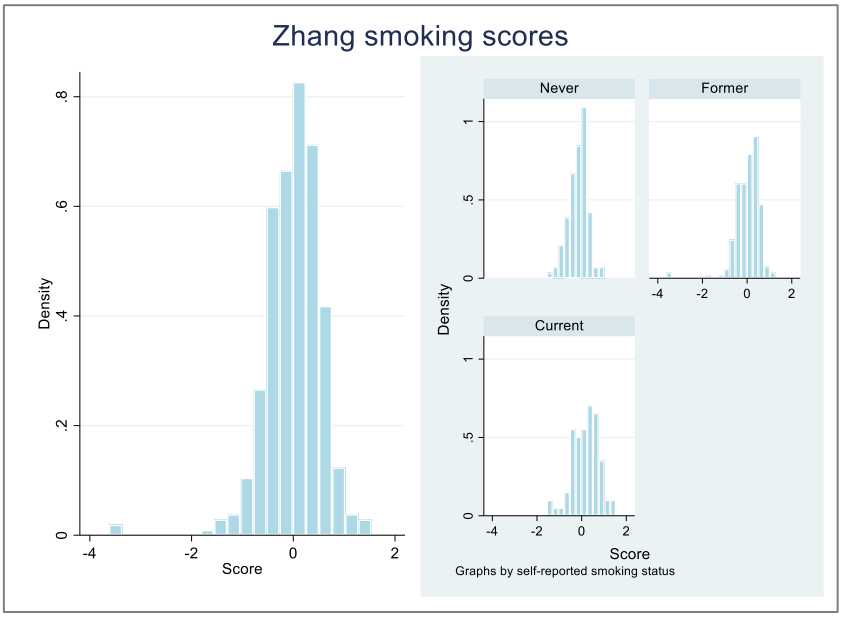
B3: Histograms showing the distribution of Joehanes (Bonferroni) DNAm scores, overall and by self-reported smoking status.



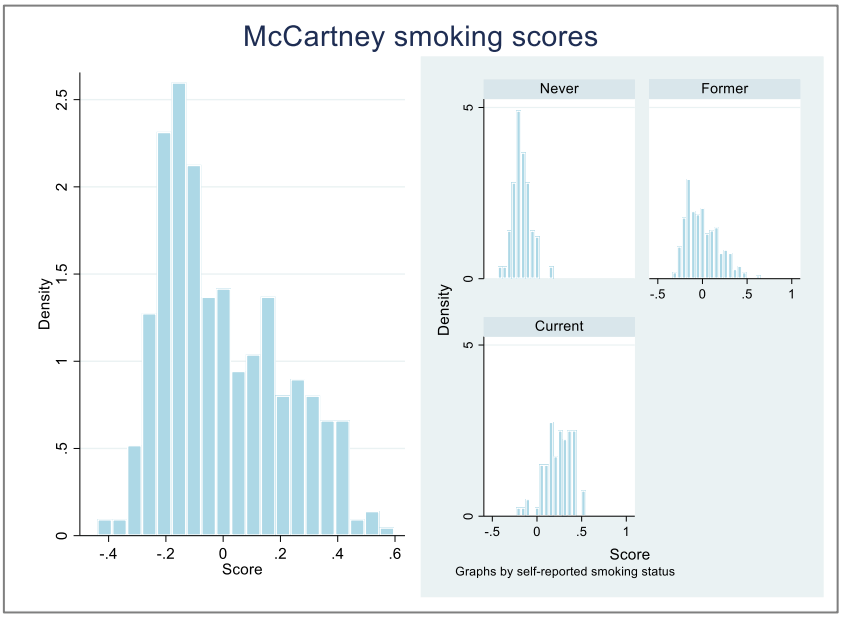
B4: Histograms showing the distribution of Joehanes (FDR) DNAm scores, overall and by self-reported smoking status.



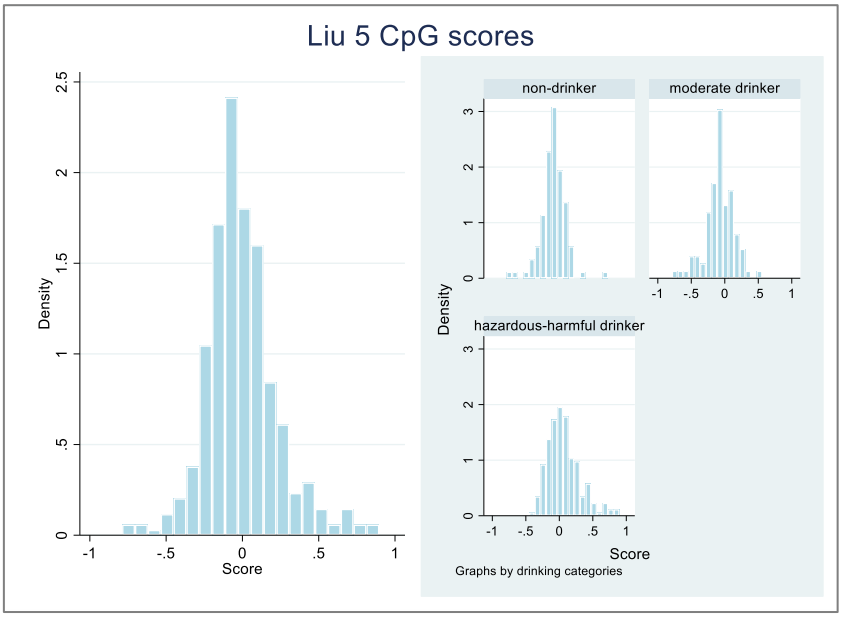
B5: Histograms showing the distribution of Zhang DNAm scores, overall and by self-reported smoking status.



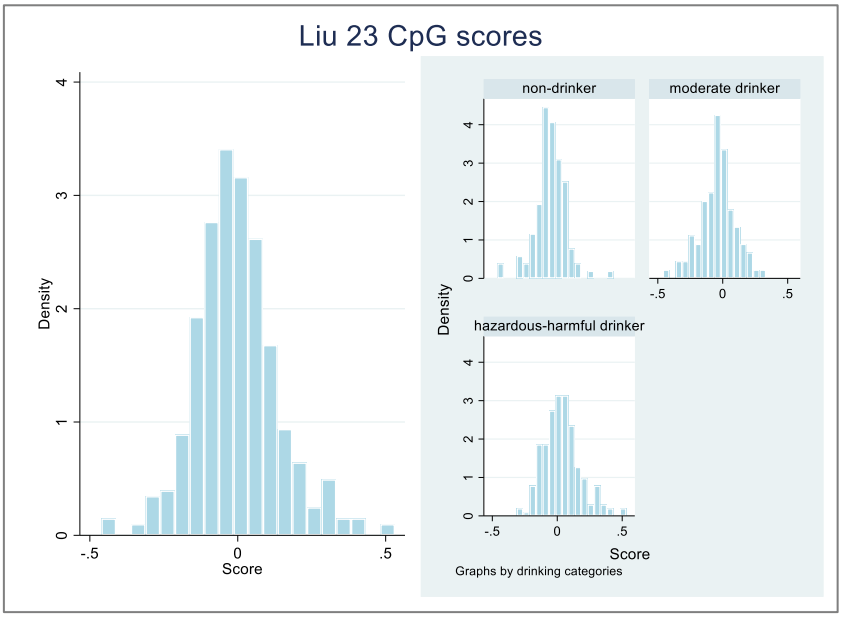
B6: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported smoking status.



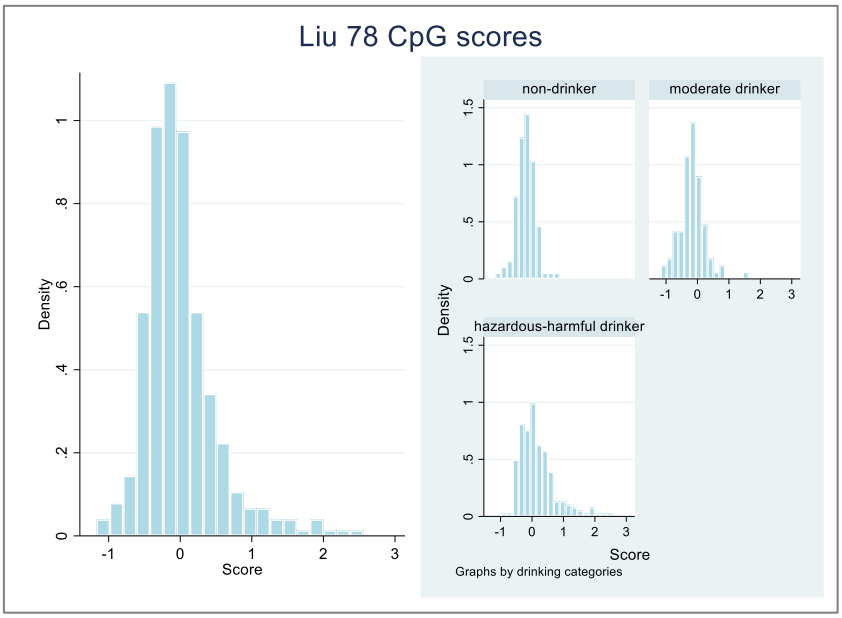
B7: Histograms showing the distribution of Lui DNAm scores (based on 5 CpGs), overall and by self-reported alcohol consumption.



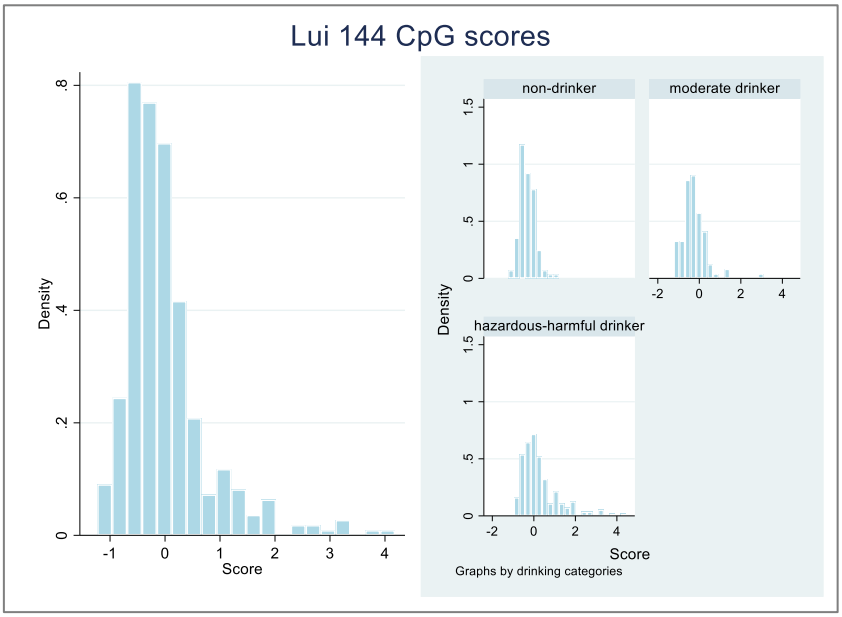
B8: Histograms showing the distribution of Lui (23 CpG) DNAm scores, overall and by self-reported alcohol consumption.



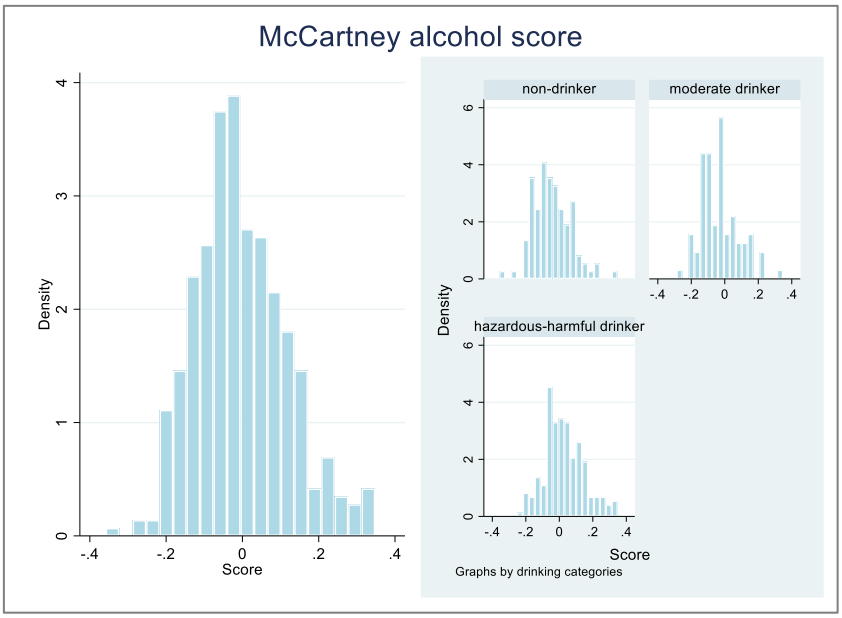
B9: Histograms showing the distribution of Lui (78 CpG) DNAm scores, overall and by self-reported alcohol consumption.



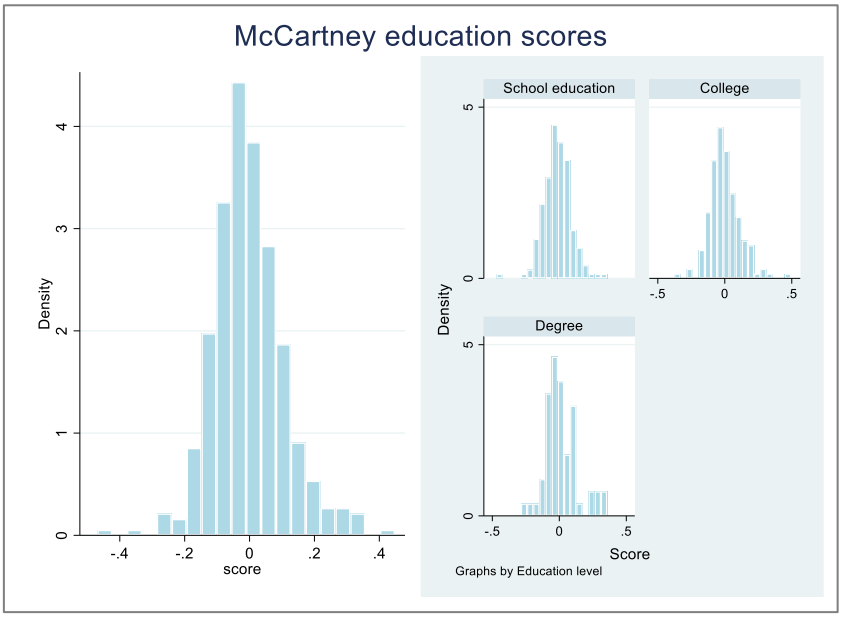
B10: Histograms showing the distribution of Lui (144 CpG) DNAm scores, overall and by self-reported alcohol consumption.



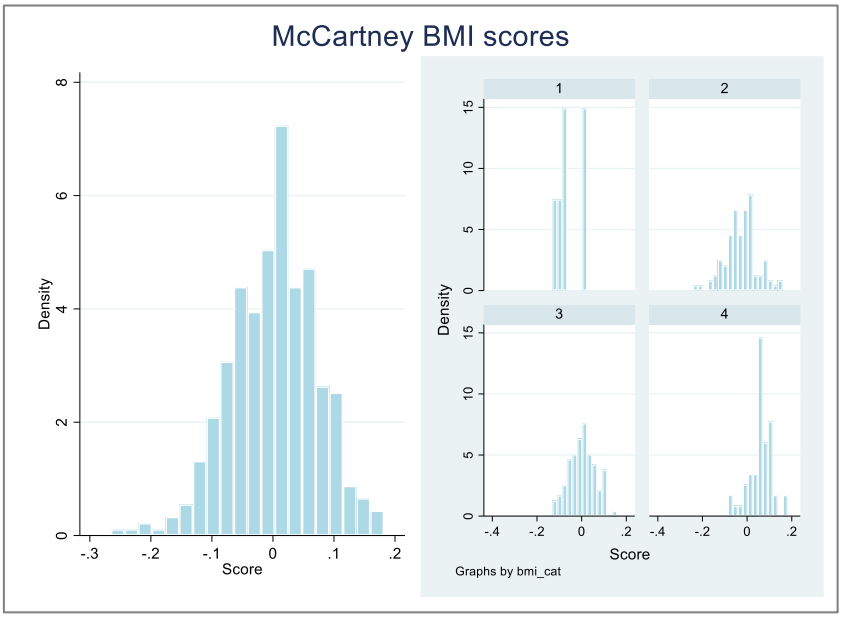
B11: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported alcohol consumption.



B12: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported educational attainment level.



B13: Histograms showing the distribution of McCartney DNAm scores, overall and by self-reported BMI.



B14: Stata code used to standardise DNAm scores, to allow direct comparison across scores with different scales.

The two smoking predictors *Joehanes* and *AHRR* have been used as a exemplars.

- **Step 1:** Obtain mean values (of DNAm score) for each category of smoking status.

```
table smoking, c(mean joehanes_scores)
```

self-reported smoking status	mean(joehan~s)
Never	-0.09904
Former	-0.02009
Current	0.192333

Positive value

```
table smoking, c(mean ahrr_scores)
```

self-reported smoking status	mean(ahrr_S~S)
Never	0.8479779
Former	0.0340207
Current	-1.249833

Negative value

- **Step 2:** Subtract the mean methylation value for current smokers from the mean methylation value for never smokers.

```
di .1923333 - -.0990415
.2913748
```

An increase in the Joehanes DNAm score of 0.29 corresponds with a change from never to current smoker.

```
di -1.249833 - 0.8479779
 -2.0978109
```

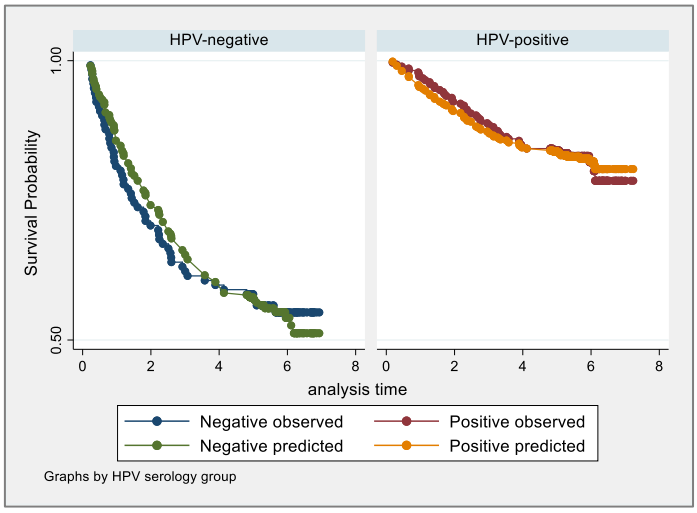
A decrease in AHRR of -2.10 corresponds with a change from never to current smoker.

- **Step 3:** Generate a new standardised variable.

```
gen joehanes_scores_stnd = joehanes_scores/.2913748
gen ahrr_scores_stnd = ahrr_scores/- 2.0978109
```

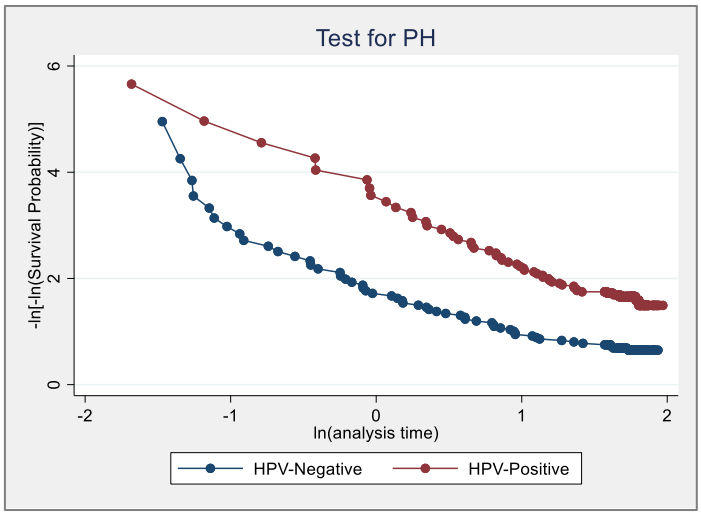
* Dividing by a negative value reverses the sign of the AHRR score. Therefore, for the standardised AHRR score, low is 'good' (i.e. never smokers have a low standardised score) and high is 'bad' (i.e. current smokers have a high standardised score). The standardised AHRR score and the standardised Johannes score are now directly comparable as they are both in the same direction and have the same units. This means that hazard ratios for the standardised scores in models for mortality are directly comparable.

B15: Kaplan-Meier observed survival curves and cox predicted survival for the variable HPV status.



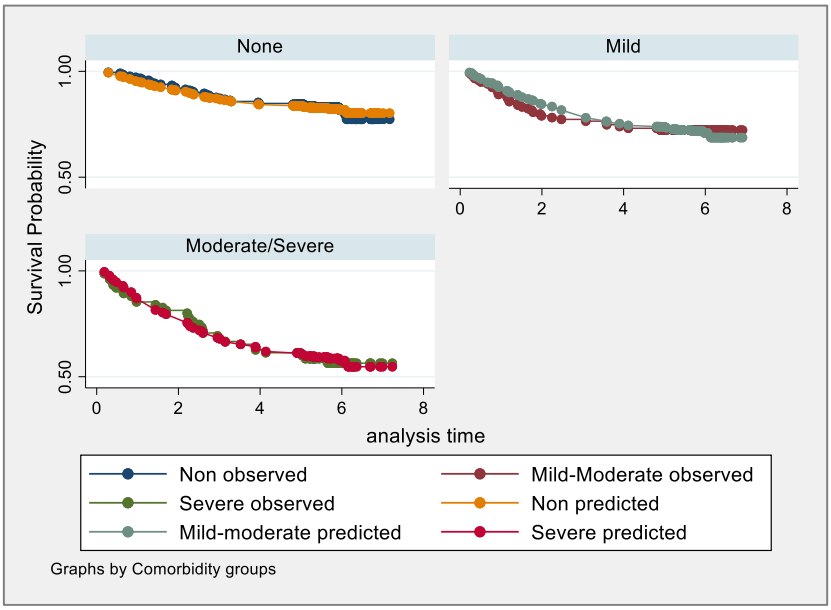
In the above plot, the closer the observed values are to the values predicted by the Cox model, the less likely the PH assumption has been violated. The plots suggest that up until approximately 3-years, there is good agreement between survival time, but after this, the PH assumption may be violated.

B16: Log-log plot testing the PH assumption for the variable HPV status.



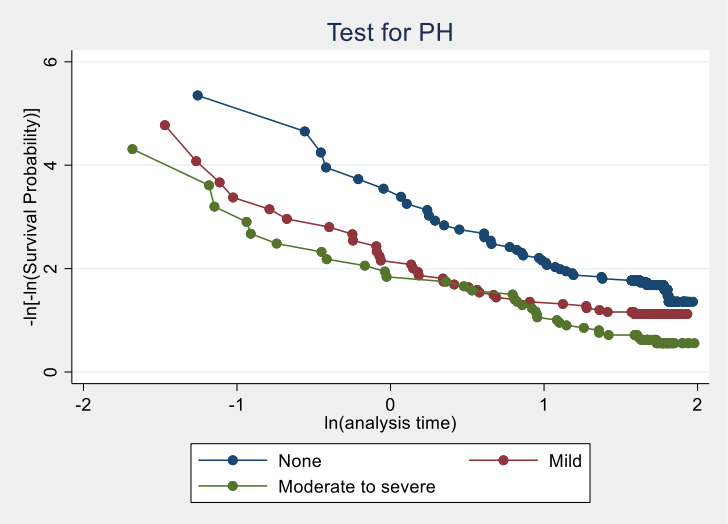
The above plot shows the ln(-ln(survival)) curves for each category HPV status (negative and positive) versus ln(time). The proportional hazards (PH) assumption is violated if the curves are not parallel. In this instance, the lines appear to converge slightly towards the upper end of the x-axis, suggesting the PH assumption may be violated. There are few events (deaths) in the HPV-positive group early on, as indicated by the small number of points.

B17: Kaplan-Meier observed survival curves and cox predicted survival for the variable comorbidity.



There are differences between the observed and predicted values in the mild and moderate to severe comorbidity groups, providing evidence of non-proportionality.

B18: Log-log plot testing the PH assumption for the variable Comorbidity.



The above plot plots −ln−ln(survival) curves for each category of comorbidity versus ln(analysis time). It provides additional evidence that the comorbidity variable may violate the PH assumption because the lines for mild and moderate to sever comorbidity cross.

B19: A comparison of the effect estimates obtained for the associations of DNAm scores with survival in the primary and complete case analyses.

	Primary analysis (n=408)				Complete case analysis (n=248)			
		95% CI				95% CI		
DNAm predictor	HR	ll	ul	*p*-value	HR	ll	ul	*p*-value
Model 1								
AHRR	2.99	1.96	4.56	3.67E-07	2.64	1.46	4.77	1.28E-03
Joehanes (FDR)	2.13	1.54	2.93	4.29E-06	2.45	1.57	3.83	7.43E-05
Joehanes (Bonferroni	2.63	1.82	3.80	2.26E-07	2.80	1.70	4.61	5.24E-05
Zhang	1.28	1.12	1.47	4.07E-04	1.23	1.02	1.49	2.86E-02
McCartney smoking	2.20	1.49	3.26	8.35E-05	2.19	1.28	3.75	4.31E-03
McCartney alcohol	1.15	1.01	1.30	2.85E-02	1.18	0.99	1.42	6.72E-02
Liu 5 CpG	1.19	1.07	1.32	1.78E-03	1.22	1.05	1.41	9.39E-03
Liu 23 CpG	1.15	1.02	1.28	1.69E-02	1.19	1.02	1.39	2.94E-02
Liu 78 CpG	1.16	1.04	1.29	7.05E-03	1.16	1.02	1.33	2.83E-02
Liu 144 CpG	1.15	1.03	1.29	1.13E-02	1.16	1.01	1.33	3.70E-02
McCartney BMI	0.88	0.77	1.01	6.25E-02	0.83	0.68	1.01	5.90E-02
McCartney education	0.94	0.89	1.00	4.57E-02	0.96	0.89	1.04	2.94E-01
Model 2								
AHRR	2.42	1.54	3.81	1.31E-04	2.51	1.32	4.75	4.82E-03
Joehanes (FDR)	1.89	1.33	2.68	3.37E-04	2.70	1.66	4.40	6.44E-05
Joehanes (Bonferroni	2.31	1.54	3.46	5.00E-05	3.21	1.83	5.64	5.13E-05
Zhang	1.22	1.07	1.39	3.65E-03	1.25	1.04	1.51	1.75E-02
McCartney smoking	1.70	1.12	2.59	1.26E-02	1.94	1.11	3.42	2.10E-02
McCartney alcohol	1.12	0.99	1.27	8.18E-02	1.21	1.02	1.45	3.32E-02
Liu 5 CpG	1.13	1.01	1.27	3.59E-02	1.28	1.08	1.52	4.26E-03
Liu 23 CpG	1.08	0.96	1.21	2.01E-01	1.23	1.03	1.47	1.94E-02
Liu 78 CpG	1.13	1.01	1.28	3.94E-02	1.22	1.05	1.43	1.09E-02
Liu 144 CpG	1.13	1.00	1.28	4.46E-02	1.21	1.04	1.42	1.59E-02
McCartney BMI	0.86	0.75	0.99	3.27E-02	0.81	0.66	0.98	3.13E-02
McCartney education	0.97	0.92	1.03	3.74E-01	0.98	0.90	1.06	5.46E-01

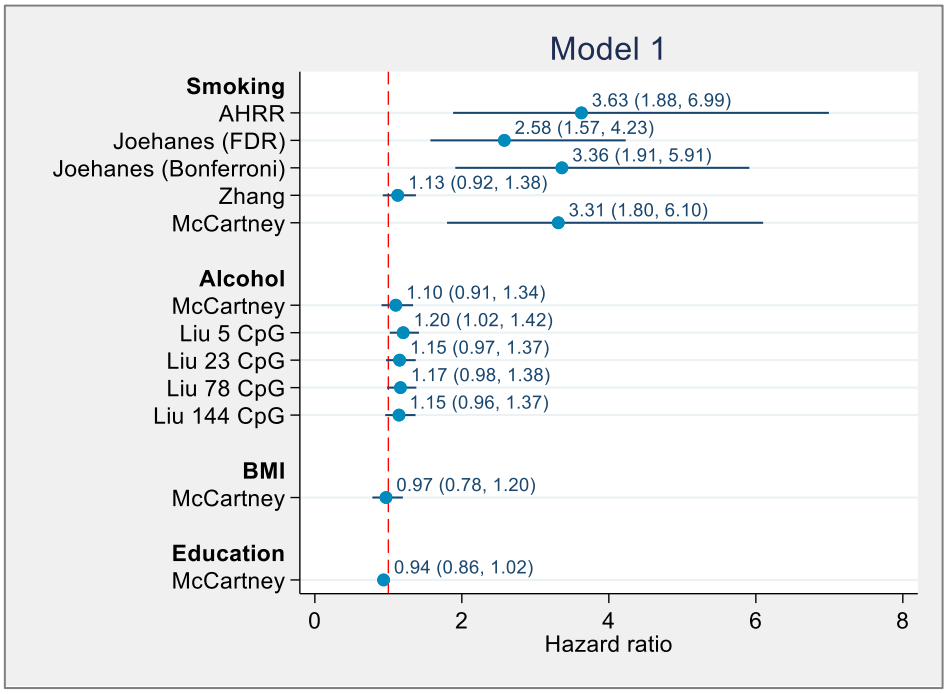
Continued on the next page.

B19 continued.

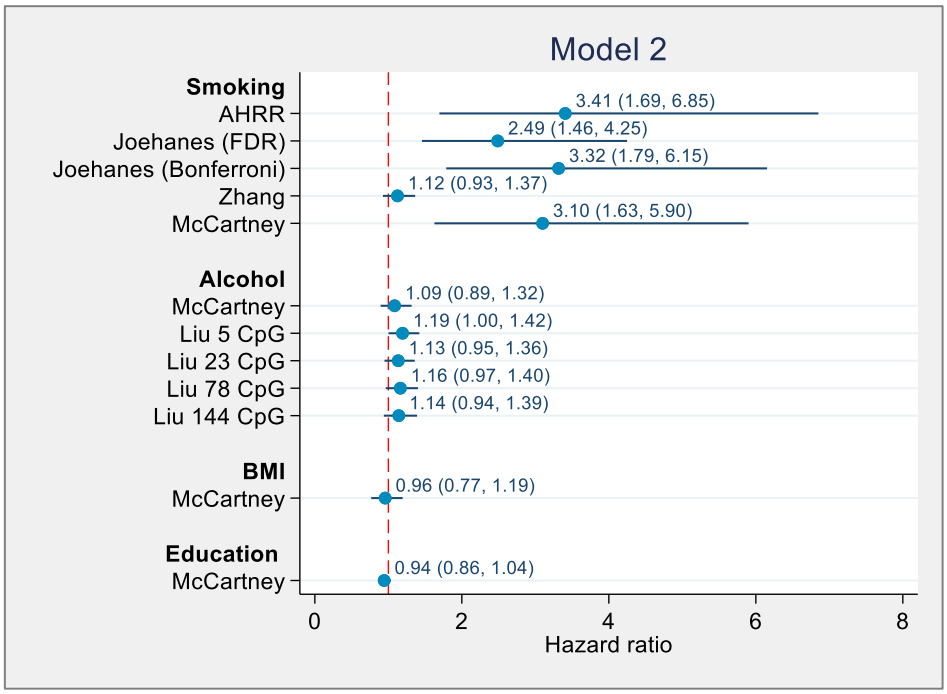
	Primary analysis (n=408)				Complete case analysis (n=248)			
		95% CI				95% CI		
DNAm predictor	HR	ll	ul	*p*-value	HR	ll	ul	*p*-value
Model 3								
AHRR	1.90	1.06	3.38	3.05E-02	1.29	0.56	2.93	5.51E-01
Joehanes (FDR)	1.56	1.04	2.34	3.12E-02	2.01	1.15	3.51	1.48E-02
Joehanes (Bonferroni	1.90	1.17	3.09	9.60E-03	2.15	1.11	4.15	2.28E-02
Zhang	1.17	1.02	1.33	2.03E-02	1.19	1.00	1.42	4.72E-02
McCartney smoking	1.17	0.69	1.97	5.55E-01	1.08	0.52	2.22	8.44E-01
McCartney alcohol	1.08	0.95	1.24	2.35E-01	1.19	0.98	1.44	8.02E-02
Liu 5 CpG	1.10	0.97	1.24	1.33E-01	1.25	1.04	1.51	1.89E-02
Liu 23 CpG	1.04	0.92	1.18	5.06E-01	1.19	0.98	1.43	7.48E-02
Liu 78 CpG	1.10	0.97	1.25	1.49E-01	1.22	1.02	1.45	2.99E-02
Liu 144 CpG	1.10	0.96	1.26	1.67E-01	1.21	1.01	1.45	4.01E-02
McCartney BMI	0.90	0.78	1.04	1.63E-01	0.84	0.68	1.03	9.28E-02
McCartney education	0.98	0.92	1.04	4.18E-01	0.98	0.90	1.07	7.01E-01
Model 4								
AHRR	1.92	1.06	3.47	3.08E-02	1.24	0.53	2.89	6.22E-01
Joehanes (FDR)	1.55	1.03	2.33	3.73E-02	1.95	1.11	3.45	2.08E-02
Joehanes (Bonferroni	1.89	1.15	3.11	1.20E-02	2.11	1.08	4.12	2.99E-02
Zhang	1.17	1.02	1.33	2.07E-02	1.19	1.00	1.42	5.02E-02
McCartney smoking	1.21	0.71	2.05	4.78E-01	1.07	0.51	2.27	8.59E-01
McCartney alcohol	1.06	0.92	1.22	4.45E-01	1.10	0.89	1.36	3.61E-01
Liu 5 CpG	1.08	0.95	1.22	2.42E-01	1.26	1.03	1.54	2.42E-02
Liu 23 CpG	1.03	0.91	1.17	6.57E-01	1.17	0.97	1.42	9.87E-02
Liu 78 CpG	1.09	0.95	1.24	2.23E-01	1.29	1.05	1.58	1.60E-02
Liu 144 CpG	1.09	0.95	1.24	2.43E-01	1.28	1.03	1.57	2.26E-02
McCartney BMI	0.92	0.79	1.07	2.76E-01	0.87	0.69	1.08	1.97E-01
McCartney education	0.99	0.93	1.06	8.39E-01	1.00	0.91	1.08	9.11E-01

Model 1: adjusted for age, gender, cell counts and batch effects; model 2: additionally adjusted for TNM stage, HPV status and comorbidity; model 3: additionally adjusted for the corresponding directly measured phenotype; model 4: additionally adjusted for the other directly measured phenotypes. Number of deaths = 105 and 56 for the primary and complete case analyses, respectively. Abbreviations: ***CI*** *confidence interval;* ***CpG****, cytosine-phosphate-guanine site;* ***FDR****, false discovery rate;* ***ll****, lower confidence level;* ***ul****, upper confidence level.*

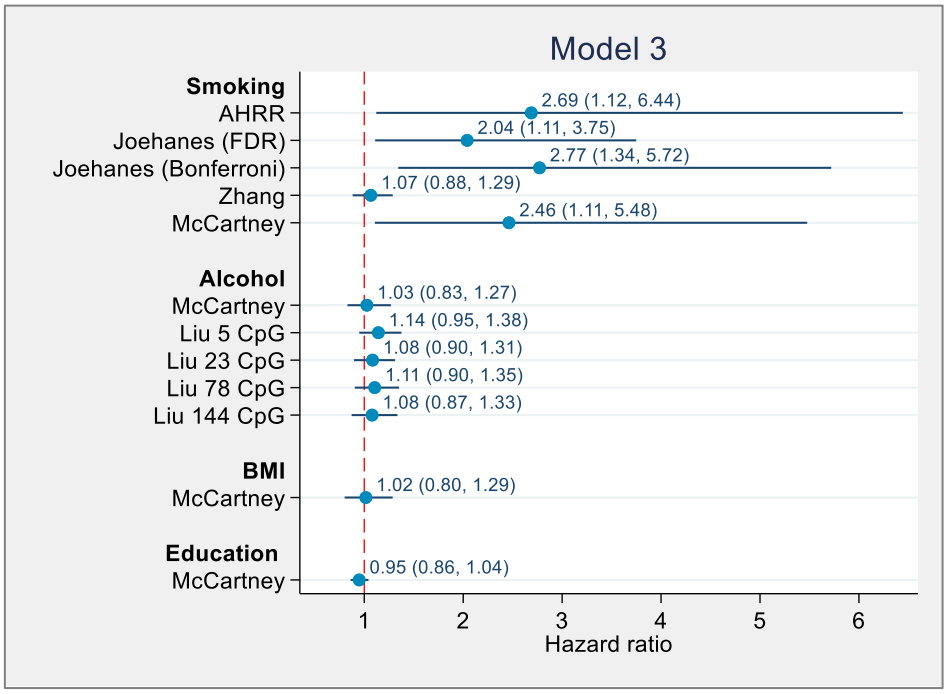
B20: Results of the sensitivity analysis, censoring data at 3-years, adjusted for age, gender, cell counts and batch effects.



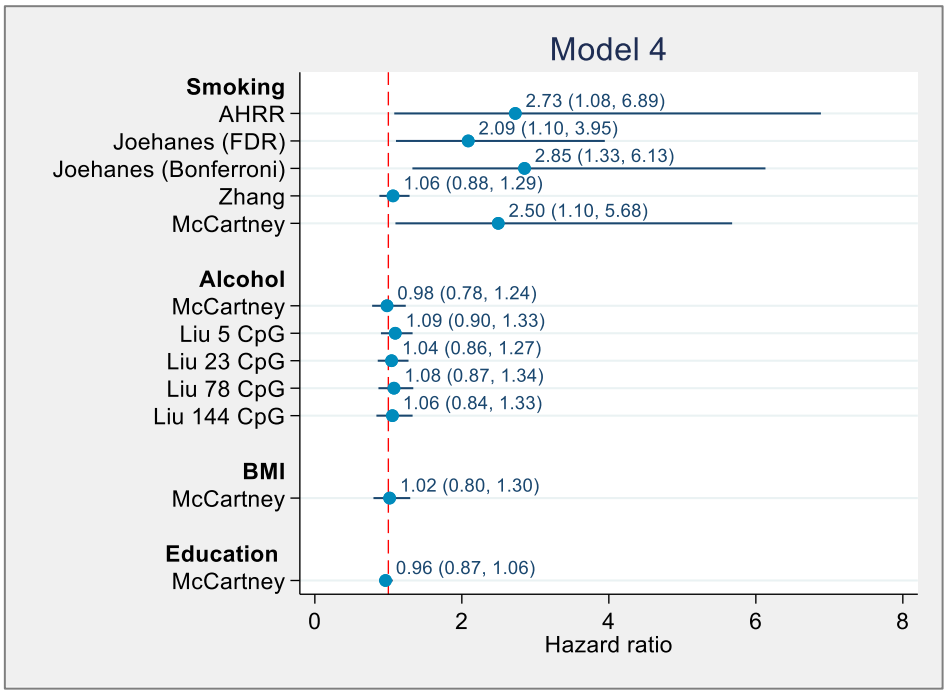
B21: Results of the sensitivity analysis, additionally adjusted for TNM stage, HPV status and comorbidity.



B22: Results of the sensitivity analysis, additionally adjusted for the corresponding directly measured phenotype.



B23: Results of the sensitivity analysis, additionally adjusted for the other directly measured phenotypes.



Appendix C

C1: Baseline characteristics of participants included in the complete case analysis.

	Dead at 3 years (n=37)		Alive at 3 years (n=188)		
Characteristic	N	%	N	%	p-value
Gender					
Male	30	81.10%	147	78.20%	
Female	7	18.90%	41	21.80%	0.695
TNM stage group					
Low	4	10.80%	30	16.00%	
High	33	89.20%	158	84.00%	0.424
HPV status					
Negative	22	59.50%	44	23.40%	
Positive	15	40.50%	144	76.60%	<0.001
Comorbidity status					
None	18	48.60%	116	61.70%	
Mild	14	37.80%	45	23.90%	
Moderate/severe	5	13.50%	27	14.40%	0.204
Smoking					
Never	2	5.40%	70	37.20%	
Former	20	54.10%	97	51.60%	
Current	15	40.50%	21	11.20%	<0.001
Alcohol					
Non-drinker	7	18.90%	52	27.70%	
Moderate	5	13.50%	49	26.10%	
Hazardous/harmful	25	67.60%	87	46.30%	0.057
Education					
School education	18	48.60%	83	44.10%	
College	14	37.80%	78	41.50%	
Degree	5	13.50%	27	14.40%	0.880-
Annual household income					
<£18,000	23	62.20%	65	34.60%	
£18000-£34,999	5	13.50%	62	33.00%	
>£35,000	9	24.30%	61	32.40%	0.005
Marital status					
single (never married)	7	18.90%	16	8.50%	
currently in relationship	16	43.20%	142	75.50%	
No longer with spouse	14	37.80%	30	16.00%	<0.001
	N	mean (SD)	N	mean (SD)	p-value
Age at baseline	37	62.76 (12.12)	188	56.93 (8.90)	0.001
Body mass index	37	23.83 (4.89)	188	26.70 (4.86)	0.001
IEAA	37	1.38 (6.52)	188	-0.44 (5.64)	0.082
EEAA	37	0.74 (3.99)	188	-0.19 (4.34)	0.226
IEAAHannum	37	1.29 (4.69)	188	-0.33 (3.84)	0.025
AgeAccelPheno	37	4.17 (5.42)	188	-1.14 (5.41)	<0.001
AgeAccelGrim	37	3.01 (7.30)	188	-0.55 (6.26)	0.002

Abbreviations: ***EEAA****, extrinsic epigenetic age acceleration;* ***IEAA****, intrinsic epigenetic age acceleration. P-value for difference based on the Chi-Square test (categorical) and one-way ANOVA (continuous). *Based on the Adult Comorbidity Evaluation-27 (ACE-27). ** values for raw epigenetic age acceleration measures.*

C2: Results of the complete case cox regression analysis (n=225).

Epigenetic clock	HR	95% confidence interval		*p*-value
		Upper	Lower	
Basic model				
IEAA	1.01	0.75	1.35	0.960
EEAA	1.41	1.04	1.91	0.026
IEAAHannum	1.36	1.04	1.80	0.027
AgeAccelGrim	1.96	1.52	2.53	2.8×10^{-07}
AgeAccelPheno	1.53	1.14	2.06	5.0×10^{-03}
Clinical model				
IEAA	0.92	0.68	1.26	0.620
EEAA	1.27	0.93	1.74	0.130
IEAAHannum	1.34	1.01	1.77	0.042
AgeAccelGrim	1.66	1.23	2.24	8.2×10^{-04}
AgeAccelPheno	1.35	0.98	1.85	0.066
Socioeconomic model				
IEAA	0.93	0.68	1.27	0.640
EEAA	1.14	0.82	1.58	0.440
IEAAHannum	1.29	0.95	1.74	0.100
AgeAccelGrim	1.45	1.04	2.02	0.027
AgeAccelPheno	1.23	0.89	1.72	0.210
Fully adjusted model				
IEAA	0.90	0.66	1.24	0.520
EEAA	1.12	0.81	1.55	0.490
IEAAHannum	1.25	0.92	1.68	0.150
AgeAccelGrim	1.24	0.83	1.85	0.290
AgeAccelPheno	1.22	0.87	1.70	0.260

Abbreviations: ***EEAA****, extrinsic epigenetic age acceleration;* ***IEAA****, intrinsic epigenetic age acceleration. Basic model adjusted for gender; clinical model additionally adjusted for tumour stage, HPV status, comorbidity and BMI; socioeconomic model additionally adjusted for education, income and marital status; Fully adjusted model adjusted for smoking status and alcohol consumption.*

C3: Estimated coefficients (uncorrected and corrected) for the clinical + AgeAccelGrim model.

	Original model			**Final model after adjustment for overfitting**		
		95% CI			95% CI	
Variable	*β*	upper	lower	*β*	upper	lower
Age	0.05	0.02	0.07	0.04	0.02	0.06
Gender						
Female	0.42	-0.14	0.99	0.35	-0.12	0.82
Tumour stage						
II	0.64	-1.56	2.85	0.53	-1.29	2.36
III	1.65	-0.38	3.69	1.37	-0.32	3.06
IV	1.85	-0.14	3.84	1.54	-0.12	3.19
HPV status						
Positive	-0.95	-1.47	-0.44	-0.79	-1.22	-0.36
Comorbidity*						
mild	0.33	-0.23	0.90	0.28	-0.19	0.75
moderate/severe	0.24	-0.38	0.85	0.20	-0.31	0.70
AgeAccelGrim	0.52	0.26	0.78	0.43	0.22	0.65

Appendix D

D1: Abbreviations, names and units of metabolic measures quantified on the Nightingale NMR platform.

Abbreviation	Full Name	Unit
	Lipoprotein Subclasses	
XXL-VLDL-P	Concentration of chylomicrons and extremely large VLDL particles	mol/l
XXL-VLDL-L	Total lipids in chylomicrons and extremely large	mmol/l
XXL-VLDL-PL	Phospholipids in chylomicrons and extremely large VLDL	mmol/l
XXL-VLDL-C	Total cholesterol in chylomicrons and extremely large VLDL	mmol/l
XXL-VLDL-CE	Cholesterol esters in chylomicrons and extremely large VLDL	mmol/l
XXL-VLDL-FC	Free cholesterol in chylomicrons and extremely large VLDL	mmol/l
XXL-VLDL-TG	Triglycerides in chylomicrons and extremely large VLDL	mmol/l
XL-VLDL-P	Concentration of very large VLDL particles	mol/l
XL-VLDL-L	Total lipids in very large VLDL	mmol/l
XL-VLDL-PL	Phospholipids in very large VLDL	mmol/l
XL-VLDL-C	Total cholesterol in very large VLDL	mmol/l
XL-VLDL-CE	Cholesterol esters in very large VLDL	mmol/l
XL-VLDL-FC	Free cholesterol in very large VLDL	mmol/l
XL-VLDL-TG	Triglycerides in very large VLDL	mmol/l
L-VLDL-P	Concentration of large VLDL particles	mol/l
L-VLDL-L	Total lipids in large VLDL	mmol/l
L-VLDL-PL	Phospholipids in large VLDL	mmol/l
L-VLDL-C	Total cholesterol in large VLDL	mmol/l
L-VLDL-CE	Cholesterol esters in large VLDL	mmol/l
L-VLDL-FC	Free cholesterol in large VLDL	mmol/l
L-VLDL-TG	Triglycerides in large VLDL	mmol/l
M-VLDL-P	Concentration of medium VLDL particles	mol/l
M-VLDL-L	Total lipids in medium VLDL	mmol/l
M-VLDL-PL	Phospholipids in medium VLDL	mmol/l
M-VLDL-C	Total cholesterol in medium VLDL	mmol/l
M-VLDL-CE	Cholesterol esters in medium VLDL	mmol/l
M-VLDL-FC	Free cholesterol in medium VLDL	mmol/l
M-VLDL-TG	Triglycerides in medium VLDL	mmol/l
S-VLDL-P	Concentration of small VLDLparticles	mol/l
S-VLDL-L	Total lipids in small VLDL	mmol/l
S-VLDL-PL	Phospholipids in small VLDL	mmol/l
S-VLDL-C	Total cholesterol in small VLDL	mmol/l
S-VLDL-CE	Cholesterol esters in small VLDL	mmol/l
S-VLDL-FC	Free cholesterol in small VLDL	mmol/l

D1 continued.

Abbreviation	Full Name	Unit
S-VLDL-TG	Triglycerides in small VLDL	mmol/l
XS-VLDL-P	Concentration of very small VLDL particles	mol/l
XS-VLDL-L	Total lipids in very small VLDL	mmol/l
XS-VLDL-PL	Phospholipids in very small VLDL	mmol/l
XS-VLDL-C	Total cholesterol in very small VLDL	mmol/l
XS-VLDL-CE	Cholesterol esters in very small VLDL	mmol/l
XS-VLDL-FC	Free cholesterol in very small VLDL	mmol/l
XS-VLDL-TG	Triglycerides in very small VLDL	mmol/l
IDL-P	Concentration of LDL particles	mol/l
IDL-L	Total lipids in LDL	mmol/l
IDL-PL	Phospholipids in LDL	mmol/l
IDL-C	Total cholesterol in LDL	mmol/l
IDL-CE	Cholesterol esters in LDL	mmol/l
IDL-FC	Free cholesterol in LDL	mmol/l
IDL-TG T	Triglycerides in LDL	mmol/l
L-LDL-P	Concentration of large LDL particles	mol/l
L-LDL-L	Total lipids in large LDL	mmol/l
L-LDL-PL	Phospholipids in large LDL	mmol/l
L-LDL-C	Total cholesterol in large LDL	mmol/l
L-LDL-CE	Cholesterol esters in large LDL	mmol/l
L-LDL-FC	Free cholesterol in large LDL	mmol/l
L-LDL-TG	Triglycerides in large LDL	mmol/l
M-LDL-P	Concentration of medium LDL particles	mol/l
M-LDL-L	Total lipids in medium LDL	mmol/l
M-LDL-PL	Phospholipids in medium LDL	mmol/l
M-LDL-C	Total cholesterol in medium LDL	mmol/l
M-LDL-CE	Cholesterol esters in medium LDL	mmol/l
M-LDL-FC	Free cholesterol in medium LDL	mmol/l
M-LDL-TG	Triglycerides in medium LDL	mmol/l
S-LDL-P	Concentration of small LDL particles	mol/l
S-LDL-L	Total lipids in small LDL	mmol/l
S-LDL-PL	Phospholipids in small LDL	mmol/l
S-LDL-C	Total cholesterol in small LDL	mmol/l
S-LDL-CE	Cholesterol esters in small LDL	mmol/l
S-LDL-FC	Free cholesterol in small LDL	mmol/l
S-LDL-TG	Triglycerides in small LDL	mmol/l

D1 continued.

Abbreviation	Full Name	Unit
XL-HDL-P	Concentration of very large HDL particles	mol/l
XL-HDL-L	Total lipids in very large HDL particles	mmol/l
XL-HDL-PL	Phospholipids in very large HDL particles	mmol/l
XL-HDL-C	Total cholesterol in very large HDL particles	mmol/l
XL-HDL-CE	Cholesterol esters in very large HDL particles	mmol/l
XL-HDL-FC	Free cholesterol in very large HDL particles	mmol/l
XL-HDL-TG	Triglycerides in very large HDL particles	mmol/l
L-HDL-P	Concentration of large HDL particles	mol/l
L-HDL-L	Total lipids in large HDL particles	mmol/l
L-HDL-PL	Phospholipids in large HDL particles	mmol/l
L-HDL-C	Total cholesterol in large HDL particles	mmol/l
L-HDL-CE	Cholesterol esters in large HDL particles	mmol/l
L-HDL-FC	Free cholesterol in large HDL particles	mmol/l
L-HDL-TG	Triglycerides in large HDL particles	mmol/l
M-HDL-P	Concentration of medium HDL particles	mol/l
M-HDL-L	Total lipids in medium HDL particles	mmol/l
M-HDL-PL	Phospholipids in medium HDL particles	mmol/l
M-HDL-C	Total cholesterol in medium HDL particles	mmol/l
M-HDL-CE	Cholesterol esters in medium HDL particles	mmol/l
M-HDL-FC	Free cholesterol in medium HDL particles	mmol/l
M-HDL-TG	Triglycerides in medium HDL particles	mmol/l
S-HDL-P	Concentration of small HDL particles	mol/l
S-HDL-L	Total lipids in small HDL particles	mmol/l
S-HDL-PL	Phospholipids in small HDL particles	mmol/l
S-HDL-C	Total cholesterol in small HDL particles	mmol/l
S-HDL-CE	Cholesterol esters in small HDL particles	mmol/l
S-HDL-FC	Free cholesterol in small HDL particles	mmol/l
S-HDL-TG	Triglycerides in small HDL particles	mmol/l
XXL-VLDL-PL_%	Phospholipids to total lipids ratio in chylomicrons and extremely large VLDL	%
XXL-VLDL-C_%	Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL	%
XXL-VLDL-CE_%	Cholesterol esters to total lipids ratio in chylomicrons and extremely large VLDL	%
XXL-VLDL-FC_%	Free cholesterol to total lipids ratio in chylomicrons and extremely large VLDL	%
XXL-VLDL-TG_%	Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL	%
XL-VLDL-PL_%	Phospholipids to total lipids ratio in very large VLDL	%
XL-VLDL-C_%	Total cholesterol to total lipids ratio in very large VLDL	%
XL-VLDL-CE_%	Cholesterol esters to total lipids ratio in very large VLDL	%

D1 continued.

Abbreviation	Full Name	Unit
XL-VLDL-FC_%	Free cholesterol to total lipids ratio in very large VLDL	%
XL-VLDL-TG_%	Triglycerides to total lipids ratio in very large VLDL	
L-VLDL-PL_%	Phospholipids to total lipids ratio in large VLDL	%
L-VLDL-C_%	Total cholesterol to total lipids ratio in large VLDL	%
L-VLDL-CE_%	Cholesterol esters to total lipids ratio in large VLDL	%
L-VLDL-FC_%	Free cholesterol to total lipids ratio in large VLDL	%
L-VLDL-TG_%	Triglycerides to total lipids ratio in large VLDL	%
M-VLDL-PL_%	Phospholipids to total lipids ratio in medium VLDL	%
M-VLDL-C_%	Total cholesterol to total lipids ratio in medium VLDL	%
M-VLDL-CE_%	Cholesterol esters to total lipids ratio in medium VLDL	%
M-VLDL-FC_%	Free cholesterol to total lipids ratio in medium VLDL	%
M-VLDL-TG_%	Triglycerides to total lipids ratio in medium VLDL	%
S-VLDL-PL_%	Phospholipids to total lipids ratio in small VLDL	%
S-VLDL-C_%	Total cholesterol to total lipids ratio in small VLDL	%
S-VLDL-CE_%	Cholesterol esters to total lipids ratio in small VLDL	%
S-VLDL-FC_%	Free cholesterol to total lipids ratio in small VLDL	%
S-VLDL-TG_%	Triglycerides to total lipids ratio in small VLDL	%
XS-VLDL-PL_%	Phospholipids to total lipids ratio in very small VLDL	%
XS-VLDL-C_%	Total cholesterol to total lipids ratio in very small VLDL	%
XS-VLDL-CE_%	Cholesterol esters to total lipids ratio in very small VLDL	%
XS-VLDL-FC_%	Free cholesterol to total lipids ratio in very small VLDL	%
XS-VLDL-TG_%	Triglycerides to total lipids ratio in very small VLDL	%
IDL-PL_%	Phospholipids to total lipids ratio in IDL	%
IDL-C_%	Total cholesterol to total lipids ratio in IDL	%
IDL-CE_%	Cholesterol esters to total lipids ratio in IDL	%
IDL-FC_%	Free cholesterol to total lipids ratio in IDL	%
IDL-TG_%	Triglycerides to total lipids ratio in IDL	%
L-LDL-PL_%	Phospholipids to total lipids ratio in large LDL	%
L-LDL-C_%	Total cholesterol to total lipids ratio in large LDL	%
L-LDL-CE_%	Cholesterol esters to total lipids ratio in large LDL	%
L-LDL-FC_%	Free cholesterol to total lipids ratio in large LDL	%
L-LDL-TG_%	Triglycerides to total lipids ratio in large LDL	%
M-LDL-PL_%	Phospholipids to total lipids ratio in medium LDL	%
M-LDL-C_%	Total cholesterol to total lipids ratio in medium LDL	%
M-LDL-CE_%	Cholesterol esters to total lipids ratio in medium LDL	%
M-LDL-FC_%	Free cholesterol to total lipids ratio in medium LDL	%

D1 continued.

Abbreviation	Full Name	Unit
M-LDL-TG_%	Triglycerides to total lipids ratio in medium LDL	%
S-LDL-PL_%	Phospholipids to total lipids ratio in small LDL	%
S-LDL-C_%	Total cholesterol to total lipids ratio in small LDL	%
S-LDL-CE_%	Cholesterol esters to total lipids ratio in small LDL	%
S-LDL-FC_%	Free cholesterol to total lipids ratio in small LDL	%
S-LDL-TG_%	Triglycerides to total lipids ratio in small LDL	%
XL-HDL-PL_%	Phospholipids to total lipids ratio in very large HDL	%
XL-HDL-C_%	Total cholesterol to total lipids ratio in very large HDL	%
XL-HDL-CE_%	Cholesterol esters to total lipids ratio in very large HDL	%
XL-HDL-FC_%	Free cholesterol to total lipids ratio in very large HDL	%
XL-HDL-TG_%	Triglycerides to total lipids ratio in very large HDL	%
L-HDL-PL_%	Phospholipids to total lipids ratio in large HDL	%
L-HDL-C_%	Total cholesterol to total lipids ratio in large HDL	%
L-HDL-CE_%	Cholesterol esters to total lipids ratio in large HDL	%
L-HDL-FC_%	Free cholesterol to total lipids ratio in large HDL	%
L-HDL-TG_%	Triglycerides to total lipids ratio in large HDL	%
M-HDL-PL_%	Phospholipids to total lipids ratio in medium HDL	%
M-HDL-C_%	Total cholesterol to total lipids ratio in medium HDL	%
M-HDL-CE_%	Cholesterol esters to total lipids ratio in medium HDL	%
M-HDL-FC_%	Free cholesterol to total lipids ratio in medium HDL	%
M-HDL-TG_%	Triglycerides to total lipids ratio in medium HDL	%
S-HDL-PL_%	Phospholipids to total lipids ratio in small HDL	%
S-HDL-C_%	Total cholesterol to total lipids ratio in small HDL	%
S-HDL-CE_%	Cholesterol esters to total lipids ratio in small HDL	%
S-HDL-FC_%	Free cholesterol to total lipids ratio in small HDL	%
S-HDL-TG_%	Triglycerides to total lipids ratio in small HDL	%
	Lipoprotein Particle Size	
VLDL-D	Mean diameter for VLDL particles	nm
LDL-D	Mean diameter for LDL particles	nm
HDL-D	Mean diameter for HDL particles	nm
	Cholesterol	
Serum-C	Serum total cholesterol	mmol/l
VLDL-C	Total cholesterol in VLDL	mmol/l
Remnant-C	Remnant cholesterol (non-HDL, non-LDL -cholesterol)	mmol/l
LDL-C	Total cholesterol in LDL	mmol/l
HDL-C	Total cholesterol in HDL	mmol/l
HDL2-C	Total cholesterol in HDL2	mmol/l

D1 continued.

Abbreviation	Full Name	Unit
HDL3-C	Total cholesterol in HDL3	mmol/l
EstC	Esterified cholesterol	mmol/l
FreeC	Free cholesterol	mmol/l
	Glycerides and Phospholipids	
Serum-TG	Serum total triglycerides	mmol/l
VLDL-TG	Triglycerides in VLDL mmol/l	mmol/l
LDL-TG	Triglycerides in LDL mmol/l	mmol/l
HDL-TG	Triglycerides in HDL mmol/l	mmol/l
TotPG	Total phosphoglycerides mmol/l	mmol/l
TG/PG	Ratio of triglycerides to phosphoglycerides	
PC	Phosphatidylcholine and other cholines	mmol/l
SM	Sphingomyelins	mmol/l
TotCho	Total cholines	mmol/l
	Apolipoproteins	
ApoA1	Apolipoprotein A-1	g/l
ApoB	Apolipoprotein B	g/l
ApoB/ApoA1	Ratio of apolipoprotein B to apolipoprotein A-I	
	Fatty Acids and Saturation Measures	
TotFA	Total fatty acids	mmol/l
DHA 22:6	docosahexaenoic acid	mmol/l
LA 18:2	linoleic acid	mmol/l
FAw3	Omega-3 fatty acids	mmol/l
FAw6	Omega-6 fatty acids	mmol/l
PUFA	Polyunsaturated fatty acids	mmol/l
MUFA	Monounsaturated fatty acids; 16:1, 18:1	mmol/l
SFA	Saturated fatty acids	
UnSat	Estimated degree of unsaturation	
	Fatty Acids (%)	
DHA/FA	Ratio of 22:6 docosahexaenoic acid to total fatty acids	%
LA/FA	Ratio of 18:2 linoleic acid to total fatty acids	%
FAw3/FA	Ratio of omega-3 fatty acids to total fatty acids	%
FAw6/FA	Ratio of omega-6 fatty acids to total fatty acids	%
PUFA/FA	Ratio of polyunsaturated fatty acids to total fatty acids	%
MUFA/FA	Ratio of monounsaturated fatty acids to total fatty acids	%
SFA/FA	Ratio of saturated fatty acids to total fatty acids	%
	Glycolysis Related Metabolites	
Glc	Glucose	mmol/l
Lac	Lactate	mmol/l
Pyr*	Pyruvate	mmol/l
Cit	Citrate	mmol/l
Glol*	Glycerol	mmol/l

D1 continued.

Abbreviation	Full name	Unit
	Amino Acids	
Ala	Alanine	mmol/l
Gln	Glutamine	mmol/l
Gly*	Glycine	mmol/l
His	Histidine	mmol/l
	Branched-chain amino acids	
Ile	Isoleucine	mmol/l
Leu	Leucine	mmol/l
Val	Valine	mmol/l
D1 continued.		
	Aromatic amino acids	
Phe	Phenylalanine	mmol/l
Tyr	Tyrosine	mmol/l
	Ketone Bodies	
Ace	Acetate	mmol/l
AcAce	Acetoacetate	mmol/l
bOHBut	3-hydroxybutyrate	mmol/l
	Fluid Balance and Inflammation	
Crea	Creatinine	mmol/l
Alb	Albumin	signal area
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein	mmol/l

**Not quantifiable for EDTA-plasma samples.*

D2: PC loadings for the metabolic trait measures.

Metabolite	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
HDL_C	-0.25	0.82	0.47	0.18	-0.01	0.06	0.00	0.04	0.02	-0.04	-0.02
ApoA1	0.15	0.81	0.53	0.12	0.00	0.11	0.00	0.09	0.03	-0.11	-0.02
HDL2_C	-0.30	0.80	0.47	0.19	0.01	0.06	-0.01	0.04	0.02	-0.03	-0.02
HDL3_C	0.28	0.80	0.36	0.13	-0.16	0.04	0.06	-0.03	0.03	-0.18	-0.03
L_HDL_PL	-0.35	0.79	0.38	0.32	0.05	0.01	-0.01	-0.01	0.00	-0.05	0.00
L_HDL_P	-0.35	0.76	0.35	0.41	0.09	0.02	-0.01	0.00	0.00	-0.02	0.00
L_HDL_L	-0.36	0.75	0.35	0.41	0.09	0.02	-0.01	0.00	0.01	-0.02	0.00
XL_HDL_PL	-0.24	0.75	0.10	0.57	0.16	-0.02	-0.02	0.02	0.04	-0.01	-0.02
L_HDL_FC	-0.39	0.73	0.32	0.43	0.11	0.02	-0.01	0.03	0.03	0.02	-0.01
S_HDL_CE	0.07	0.72	0.12	-0.54	-0.34	0.03	0.06	-0.02	0.00	-0.16	-0.02
L_HDL_C	-0.39	0.72	0.32	0.45	0.13	0.03	-0.01	0.02	0.02	0.01	-0.01
XL_HDL_P	-0.06	0.72	0.07	0.65	0.20	0.01	-0.03	0.05	0.05	-0.02	-0.03
XL_HDL_L	-0.06	0.72	0.08	0.65	0.20	0.02	-0.03	0.05	0.05	-0.02	-0.03
L_HDL_CE	-0.39	0.71	0.32	0.46	0.13	0.03	-0.01	0.02	0.01	0.01	-0.01
IDL_FC	0.65	0.70	-0.30	-0.05	-0.03	0.02	-0.01	0.02	-0.03	-0.23	0.01
L_LDL_FC	0.65	0.69	-0.30	-0.10	-0.04	0.03	0.00	0.04	-0.01	-0.22	0.00
XL_HDL_C	0.10	0.66	0.05	0.67	0.22	0.06	-0.03	0.09	0.07	-0.01	-0.04
EstC	0.72	0.66	-0.08	-0.03	-0.10	0.09	0.02	0.11	0.00	-0.20	-0.02
S_HDL_C	0.05	0.66	0.31	-0.54	-0.36	0.03	0.05	-0.03	-0.01	-0.17	-0.01
XL_HDL_CE	0.15	0.66	0.00	0.66	0.22	0.07	-0.02	0.10	0.07	-0.01	-0.04
XL_HDL_FC	-0.03	0.65	0.16	0.68	0.20	0.03	-0.04	0.07	0.07	0.00	-0.05
Serum_C	0.75	0.65	-0.05	-0.04	-0.04	0.06	0.00	0.06	0.00	-0.23	-0.01
IDL_PL	0.72	0.62	-0.29	-0.05	-0.04	0.00	-0.01	0.01	-0.03	-0.25	0.01
S_LDL_CE	0.69	0.62	-0.28	-0.22	-0.07	0.06	0.02	0.06	0.01	-0.20	-0.01
S_LDL_C	0.71	0.61	-0.24	-0.21	-0.07	0.07	0.02	0.07	0.01	-0.19	-0.01
M_HDL_FC	-0.23	0.60	0.72	-0.17	-0.17	0.03	-0.01	-0.01	-0.02	-0.11	0.00
L_LDL_PL	0.76	0.60	-0.22	-0.13	-0.05	0.03	-0.01	0.04	-0.01	-0.23	0.00
LDL_C	0.73	0.60	-0.26	-0.17	-0.05	0.05	0.01	0.05	0.00	-0.21	-0.01
L_LDL_C	0.74	0.60	-0.26	-0.14	-0.04	0.04	0.00	0.05	-0.01	-0.22	0.00
M_LDL_CE	0.72	0.59	-0.28	-0.21	-0.06	0.05	0.01	0.06	0.00	-0.20	-0.01
M_LDL_C	0.73	0.59	-0.27	-0.20	-0.06	0.05	0.01	0.06	0.01	-0.20	-0.01
L_LDL_L	0.77	0.58	-0.24	-0.10	-0.05	0.02	0.00	0.03	-0.01	-0.24	0.00
FreeC	0.77	0.58	0.04	-0.07	0.13	-0.04	-0.07	-0.08	0.01	-0.29	0.02
M_LDL_FC	0.76	0.58	-0.18	-0.15	-0.08	0.08	0.02	0.07	0.02	-0.20	-0.02
IDL_C	0.77	0.57	-0.26	-0.07	-0.02	0.05	-0.01	0.05	-0.02	-0.23	0.00
TotCho	0.70	0.57	0.24	0.12	0.06	0.02	-0.03	-0.09	-0.05	-0.30	0.05
S_LDL_FC	0.77	0.56	-0.11	-0.17	-0.09	0.11	0.03	0.09	0.03	-0.18	-0.02
S_LDL_L	0.78	0.56	-0.16	-0.17	-0.08	0.06	0.02	0.05	0.01	-0.21	-0.02
L_LDL_CE	0.77	0.56	-0.25	-0.15	-0.04	0.04	0.00	0.05	0.00	-0.22	0.00
L_LDL_P	0.79	0.56	-0.23	-0.08	-0.05	0.02	0.00	0.02	-0.01	-0.24	0.00
M_LDL_L	0.78	0.56	-0.22	-0.15	-0.07	0.04	0.01	0.04	0.00	-0.22	-0.01
S_LDL_P	0.79	0.55	-0.15	-0.16	-0.08	0.05	0.01	0.05	0.01	-0.22	-0.01
M_LDL_P	0.79	0.55	-0.22	-0.14	-0.07	0.03	0.01	0.03	0.00	-0.23	-0.01

D2 continued.

Metabolite	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
IDL_L	0.80	0.54	-0.25	-0.02	-0.03	0.01	-0.01	0.01	-0.03	-0.25	0.01
pc	0.75	0.54	0.12	0.21	-0.13	0.05	0.01	-0.01	0.00	-0.25	0.02
M_HDL_PL	-0.14	0.53	0.80	-0.13	-0.17	-0.01	-0.02	-0.03	-0.04	-0.15	-0.01
M_HDL_CE	-0.31	0.53	0.72	-0.24	-0.12	0.09	0.01	0.04	0.00	-0.03	-0.01
M_HDL_L	-0.18	0.53	0.78	-0.18	-0.15	0.03	-0.01	0.00	-0.03	-0.11	0.00
UnSat	-0.37	0.53	-0.20	-0.22	-0.01	-0.15	0.30	0.19	0.24	0.24	-0.03
M_HDL_P	-0.16	0.52	0.80	-0.18	-0.16	0.02	-0.02	-0.01	-0.03	-0.12	0.00
IDL_P	0.82	0.51	-0.25	0.00	-0.04	0.00	-0.01	0.00	-0.03	-0.26	0.01
IDL_CE	0.81	0.51	-0.24	-0.07	-0.01	0.06	-0.02	0.06	-0.02	-0.22	0.00
TotPG	0.71	0.51	0.28	0.22	-0.04	0.05	-0.03	-0.05	-0.06	-0.28	0.04
S_LDL_PL	0.84	0.50	0.02	-0.15	-0.10	0.08	0.01	0.07	0.02	-0.22	-0.02
L_HDL_TG	0.14	0.48	0.14	0.69	0.17	0.01	-0.05	-0.09	-0.06	-0.14	0.06
XS_VLDL_PL	0.81	0.47	-0.28	-0.04	-0.08	-0.06	0.00	-0.04	-0.05	-0.29	0.02
M_LDL_PL	0.86	0.47	-0.10	-0.14	-0.08	0.04	0.00	0.05	0.01	-0.23	-0.02
FAw6_FA	-0.06	0.44	0.06	-0.56	0.63	0.03	-0.12	-0.16	0.09	-0.04	-0.03
PUFA_FA	-0.03	0.43	0.09	-0.57	0.66	-0.06	-0.03	-0.09	0.05	-0.03	-0.01
XS_VLDL_FC	0.84	0.42	-0.31	-0.06	-0.09	-0.04	0.01	-0.05	-0.06	-0.29	0.03
FAw6	0.69	0.40	0.19	-0.31	0.41	-0.03	-0.08	-0.16	0.04	-0.25	-0.01
pufa	0.70	0.39	0.21	-0.32	0.42	-0.08	-0.03	-0.12	0.02	-0.24	0.00
sm	0.31	0.36	0.27	-0.34	0.48	-0.19	-0.14	-0.24	-0.03	-0.26	0.04
la	0.63	0.33	0.19	-0.37	0.48	0.07	-0.10	-0.13	0.04	-0.19	-0.03
XS_VLDL_C	0.84	0.32	-0.40	0.02	-0.06	-0.03	0.00	-0.04	-0.06	-0.26	0.03
LA_FA	0.09	0.32	0.07	-0.55	0.64	0.14	-0.13	-0.12	0.06	-0.04	-0.05
M_LDL_TG	0.80	0.29	0.01	0.37	-0.14	-0.17	-0.01	-0.21	-0.06	-0.36	0.02
dha	0.40	0.28	0.15	0.08	0.02	-0.67	0.45	0.13	-0.02	-0.10	0.02
XS_VLDL_CE	0.83	0.28	-0.44	0.05	-0.05	-0.02	0.00	-0.04	-0.06	-0.25	0.03
S_HDL_L	0.04	0.27	0.75	-0.46	-0.35	-0.01	0.01	-0.07	-0.06	-0.19	0.00
Alb	0.24	0.27	0.10	0.00	-0.36	0.13	0.27	-0.02	0.46	0.16	-0.02
S_HDL_P	0.07	0.24	0.76	-0.45	-0.35	-0.02	0.01	-0.07	-0.07	-0.20	0.00
XS_VLDL_L	0.91	0.24	-0.28	0.02	-0.08	-0.06	0.00	-0.07	-0.05	-0.29	0.02
L_LDL_TG	0.83	0.21	-0.05	0.37	-0.12	-0.16	-0.02	-0.19	-0.06	-0.35	0.02
LDL_TG	0.85	0.20	0.00	0.35	-0.12	-0.16	-0.01	-0.19	-0.06	-0.35	0.02
FAw3	0.51	0.20	0.27	-0.25	0.39	-0.41	0.32	0.19	-0.12	-0.09	0.08
XS_VLDL_P	0.93	0.19	-0.26	0.04	-0.08	-0.07	-0.01	-0.08	-0.05	-0.29	0.02
DHA_FA	-0.20	0.19	0.05	0.09	-0.02	-0.72	0.53	0.23	0.00	0.10	-0.01
FAw3_FA	0.11	0.16	0.20	-0.34	0.47	-0.47	0.40	0.26	-0.14	0.04	0.09
Remnant_C	0.98	0.14	-0.14	-0.03	-0.01	0.02	-0.01	0.03	-0.01	-0.23	0.01
ApoB	0.98	0.14	-0.09	-0.09	-0.02	0.03	-0.01	0.05	0.01	-0.21	0.00
TotFA	0.95	0.13	0.21	0.05	0.06	-0.05	-0.02	-0.09	-0.01	-0.28	0.02
S_VLDL_CE	0.93	0.09	-0.27	-0.11	-0.06	-0.03	0.01	-0.03	-0.03	-0.24	0.03
S_HDL_FC	-0.07	0.07	0.91	-0.27	-0.25	0.01	0.01	-0.05	-0.05	-0.11	0.00
S_LDL_TG	0.93	0.05	0.16	0.22	-0.10	-0.11	-0.02	-0.12	-0.03	-0.31	0.01
bOHBut	-0.07	0.05	0.00	0.15	-0.10	-0.55	0.26	-0.10	0.05	-0.05	-0.17

D2 continued.

Metabolite	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
IDL_TG	0.87	0.02	-0.09	0.33	-0.11	-0.17	-0.02	-0.20	-0.06	-0.33	0.02
XL_HDL_TG	0.72	0.02	-0.01	0.60	0.23	0.01	-0.04	0.02	-0.01	-0.12	0.02
Ace	0.01	0.00	0.10	0.10	-0.04	0.05	-0.03	0.05	0.00	0.01	0.86
His	0.12	-0.02	0.06	-0.12	-0.04	0.31	0.33	0.06	0.56	0.39	0.18
S_HDL_PL	-0.11	-0.03	0.90	-0.30	-0.23	-0.02	-0.02	-0.05	-0.09	-0.12	0.01
S_VLDL_C	0.97	-0.03	-0.16	-0.10	-0.07	-0.04	0.00	-0.04	-0.02	-0.24	0.03
mufa	0.86	-0.04	0.13	0.30	-0.24	-0.13	-0.04	-0.16	0.08	-0.27	0.04
Tyr	0.05	-0.05	0.07	0.18	0.18	0.45	0.43	-0.13	-0.31	0.04	0.14
Phe	0.03	-0.06	-0.19	0.22	0.08	0.43	0.51	-0.34	-0.10	0.08	0.03
Gln	-0.12	-0.07	0.12	-0.22	0.27	-0.04	-0.01	-0.03	0.07	0.06	0.28
VLDL_C	0.98	-0.17	-0.04	0.00	-0.01	0.00	-0.01	0.01	-0.01	-0.19	0.01
Val	0.23	-0.17	0.04	0.00	0.07	0.34	0.74	-0.17	-0.04	0.17	-0.04
HDL_TG	0.82	-0.18	0.25	0.42	0.03	-0.12	-0.06	-0.11	-0.05	-0.26	0.03
M_HDL_TG	0.72	-0.22	0.46	-0.02	-0.09	-0.17	-0.07	-0.11	-0.10	-0.30	0.03
Crea	0.13	-0.22	-0.06	-0.15	0.04	0.13	0.21	-0.22	-0.07	-0.02	-0.17
S_VLDL_FC	0.96	-0.24	0.07	-0.06	-0.08	-0.06	-0.01	-0.05	0.00	-0.23	0.01
Leu	0.49	-0.26	0.15	0.13	0.09	0.27	0.65	-0.15	0.03	0.12	-0.04
M_VLDL_CE	0.95	-0.26	0.02	0.03	0.04	0.03	-0.02	0.05	0.01	-0.15	0.01
Gp	0.58	-0.28	0.02	0.22	0.02	-0.10	0.05	-0.27	0.04	-0.17	-0.08
S_VLDL_PL	0.93	-0.28	0.11	-0.09	-0.10	-0.08	-0.02	-0.06	0.00	-0.23	0.00
XS_VLDL_TG	0.92	-0.29	0.02	0.11	-0.09	-0.12	-0.02	-0.13	-0.02	-0.26	0.02
MUFA_FA	0.08	-0.30	-0.11	0.53	-0.65	-0.18	-0.04	-0.14	0.19	-0.04	0.04
S_VLDL_L	0.94	-0.31	0.04	-0.08	-0.05	-0.05	-0.01	-0.02	0.01	-0.19	0.01
SFA_FA	-0.08	-0.33	0.01	0.21	-0.16	0.45	0.14	0.45	-0.44	0.14	-0.05
S_VLDL_P	0.93	-0.34	0.06	-0.07	-0.04	-0.05	-0.01	-0.02	0.02	-0.18	0.00
XXL_VLDL_CE	0.90	-0.34	0.10	0.14	0.10	0.08	-0.03	0.10	-0.02	-0.11	-0.01
M_VLDL_C	0.92	-0.35	0.11	0.02	0.03	0.02	-0.02	0.05	0.02	-0.13	0.00
XXL_VLDL_C	0.87	-0.39	0.18	0.13	0.09	0.07	-0.03	0.10	0.00	-0.09	-0.01
Ile	0.61	-0.40	0.19	0.10	0.15	0.20	0.48	-0.11	0.04	0.08	-0.04
Serum_TG	0.88	-0.42	0.19	0.06	0.00	-0.03	-0.02	0.00	0.02	-0.15	-0.01
XXL_VLDL_PL	0.84	-0.42	0.29	0.08	0.08	0.07	-0.03	0.10	0.02	-0.08	-0.03
XXL_VLDL_L	0.84	-0.43	0.25	0.08	0.09	0.07	-0.03	0.11	0.02	-0.07	-0.03
XXL_VLDL_P	0.84	-0.43	0.26	0.08	0.10	0.07	-0.03	0.11	0.02	-0.07	-0.03
M_VLDL_PL	0.88	-0.44	0.17	-0.02	0.02	0.01	-0.02	0.04	0.03	-0.11	-0.01
XL_VLDL_FC	0.84	-0.44	0.25	0.10	0.07	0.06	-0.02	0.09	0.02	-0.08	-0.02
XXL_VLDL_TG	0.83	-0.44	0.27	0.07	0.10	0.08	-0.03	0.11	0.02	-0.07	-0.03
XL_VLDL_PL	0.85	-0.44	0.26	0.07	0.06	0.04	-0.03	0.08	0.03	-0.09	-0.02
XL_VLDL_C	0.85	-0.45	0.21	0.11	0.08	0.05	-0.02	0.09	0.02	-0.08	-0.02
M_VLDL_FC	0.87	-0.45	0.20	0.01	0.03	0.01	-0.02	0.04	0.03	-0.11	-0.01
XXL_VLDL_FC	0.82	-0.46	0.29	0.11	0.08	0.06	-0.03	0.09	0.03	-0.07	-0.02
L_VLDL_CE	0.87	-0.46	0.15	0.05	0.05	0.02	-0.03	0.05	0.02	-0.11	-0.01
XL_VLDL_CE	0.85	-0.46	0.18	0.12	0.08	0.05	-0.03	0.08	0.01	-0.08	-0.01
M_VLDL_L	0.86	-0.46	0.17	-0.02	0.03	0.01	-0.02	0.05	0.04	-0.10	-0.01

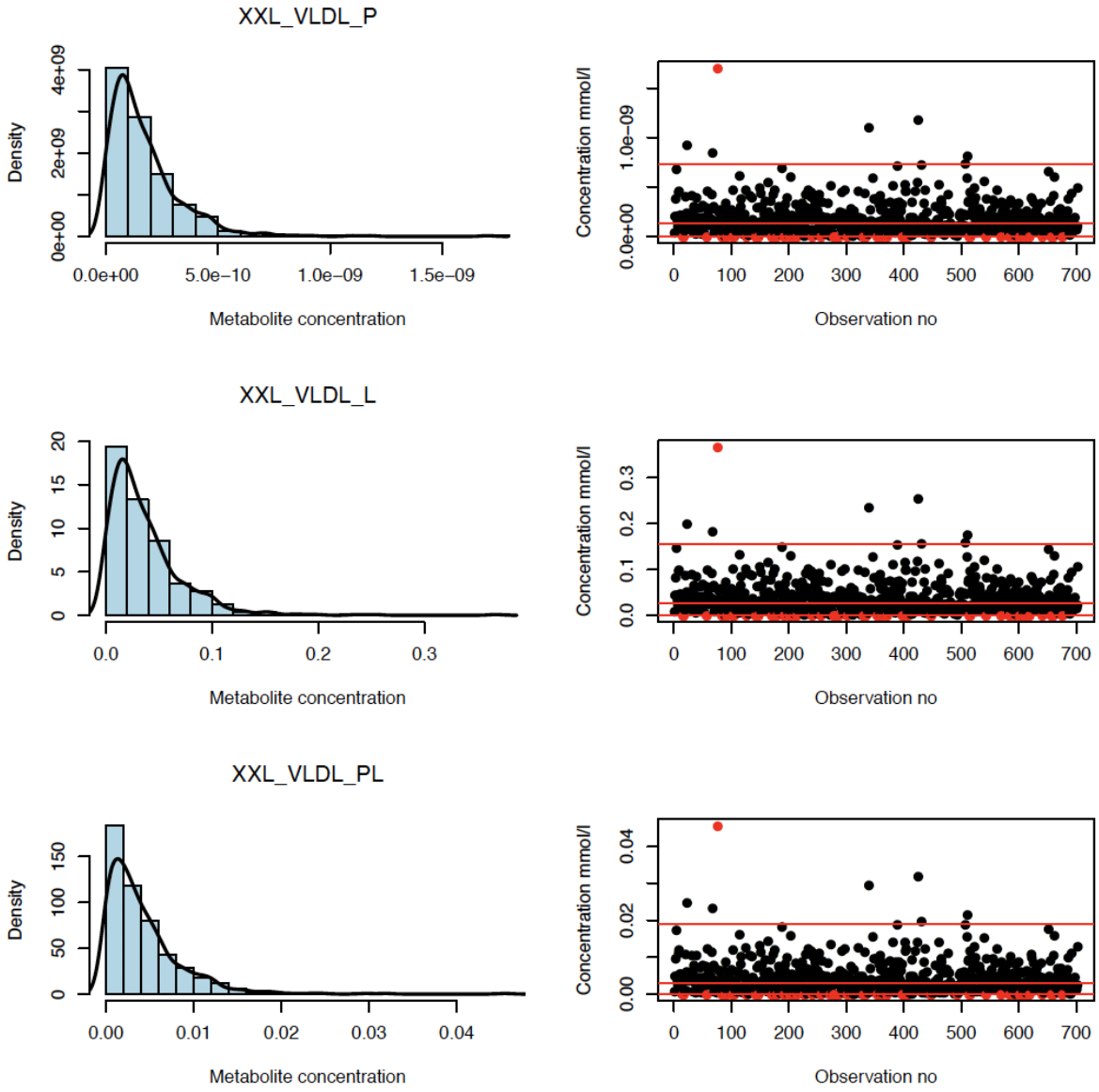
D2 continued.

Metabolite	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
L_VLDL_PL	0.85	-0.47	0.24	0.02	0.03	0.01	-0.02	0.04	0.04	-0.11	-0.01
L_VLDL_FC	0.83	-0.47	0.27	0.05	0.04	0.03	-0.02	0.06	0.04	-0.09	-0.02
M_VLDL_P	0.86	-0.48	0.18	-0.03	0.03	0.01	-0.02	0.05	0.04	-0.10	-0.01
XL_VLDL_L	0.82	-0.48	0.26	0.07	0.06	0.04	-0.03	0.08	0.04	-0.08	-0.02
XL_VLDL_P	0.82	-0.49	0.27	0.06	0.06	0.03	-0.03	0.07	0.04	-0.08	-0.02
L_VLDL_L	0.83	-0.49	0.24	0.02	0.04	0.01	-0.02	0.05	0.04	-0.09	-0.01
L_VLDL_P	0.83	-0.49	0.24	0.01	0.04	0.01	-0.02	0.05	0.04	-0.09	-0.01
S_VLDL_TG	0.84	-0.50	0.16	-0.05	-0.01	-0.03	-0.02	0.00	0.04	-0.12	-0.01
VLDL_TG	0.84	-0.50	0.21	-0.01	0.02	0.00	-0.02	0.04	0.04	-0.10	-0.01
XL_VLDL_TG	0.80	-0.50	0.28	0.05	0.06	0.03	-0.03	0.07	0.04	-0.07	-0.02
L_VLDL_TG	0.81	-0.51	0.26	0.00	0.04	0.01	-0.02	0.05	0.04	-0.09	-0.02
S_HDL_TG	0.76	-0.52	0.17	0.14	-0.11	-0.15	-0.03	-0.15	-0.04	-0.22	0.00
M_VLDL_TG	0.82	-0.52	0.20	-0.05	0.03	0.01	-0.02	0.05	0.05	-0.08	-0.01

D3: Proportion of missing data in the metabolomics dataset (n=1,483).

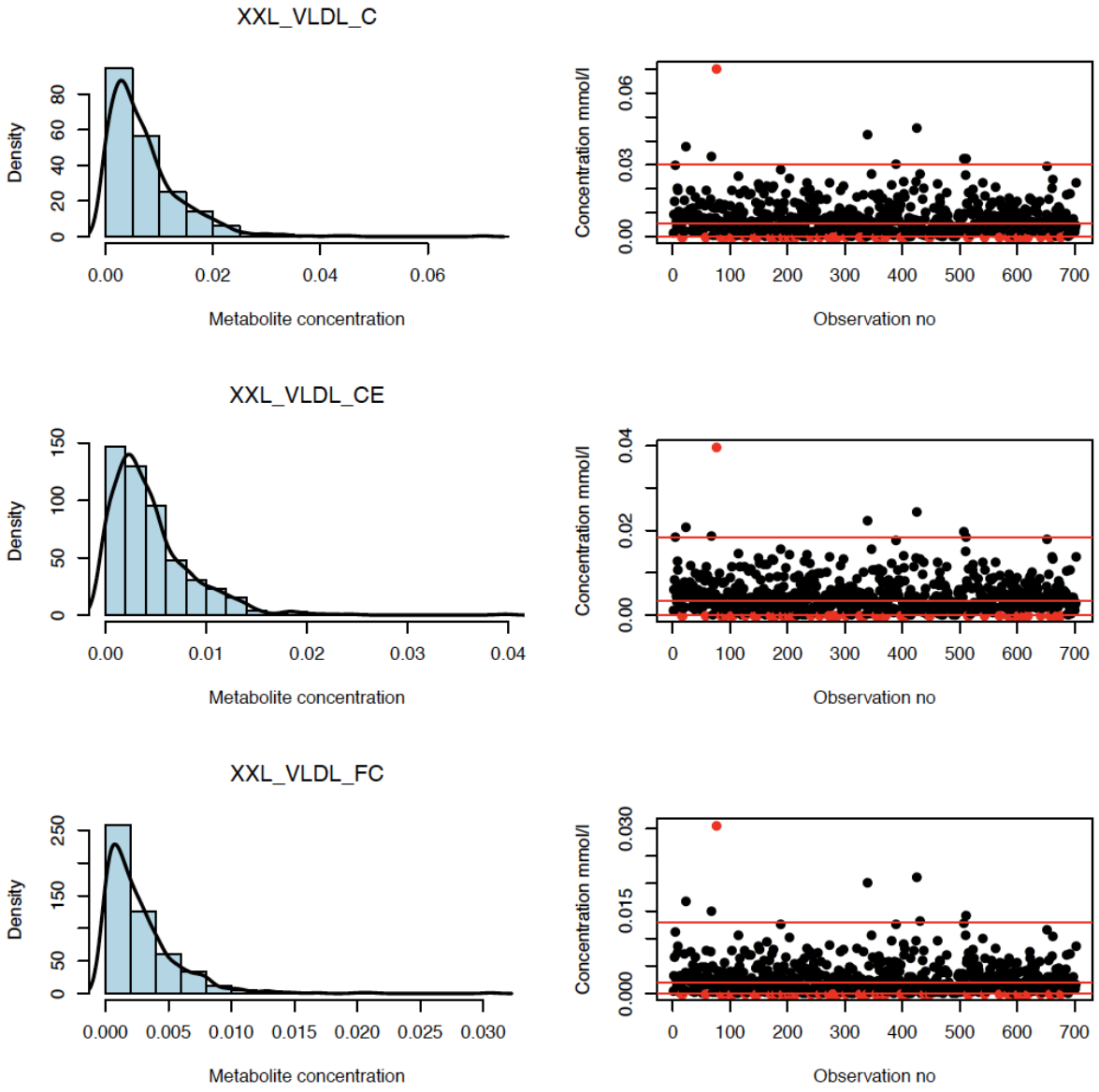
Variable	N	% Missing
Gender	0	0
Age	10	0.67
Ethnicity	35	2.36
BMI	603	40.66
hn1_ICD_gr~f	0	0
TNM stage	2	0.13
HPV status	4	0.27
Comorbidity	24	1.62
Treatment group	0	0
Annual household income	479	32.3
IMD group	372	25.08
Education	394	26.57
Marital status	359	24.21
Smoking status	399	26.9
Alcohol consumption	375	25.29

D4: Density histograms and scatter plots showing the observed data distributions of metabolic trait measures (n=703).

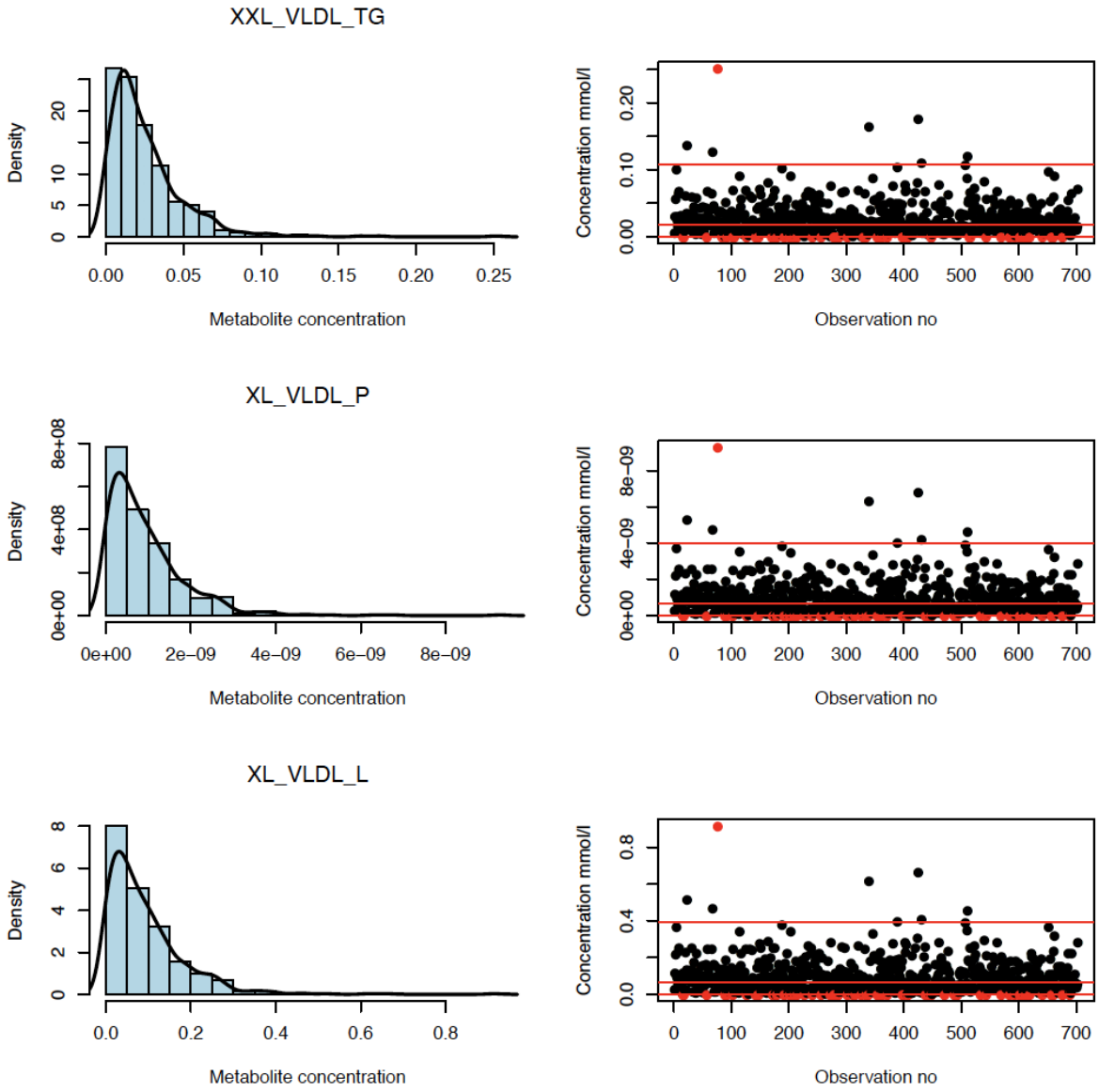


Only those metabolic trait measures included in Chapter 10 are presented. Each dot on the right hand scatter plot marks a single observation (metabolite measure). Horizontal red lines represent the 99th (top line), 50th (middle line) and 1st (bottom line) percentiles. Red dots mark are those observations that sit in the upper and lower 0.1% of the distribution.

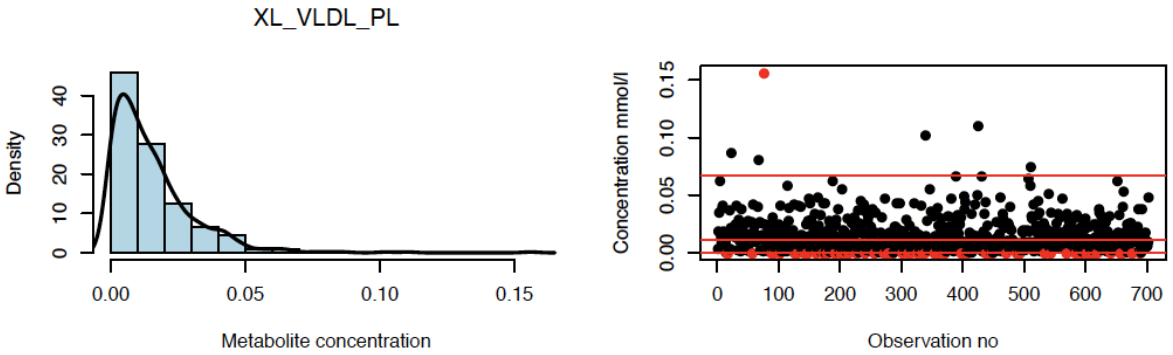
D4 continued.

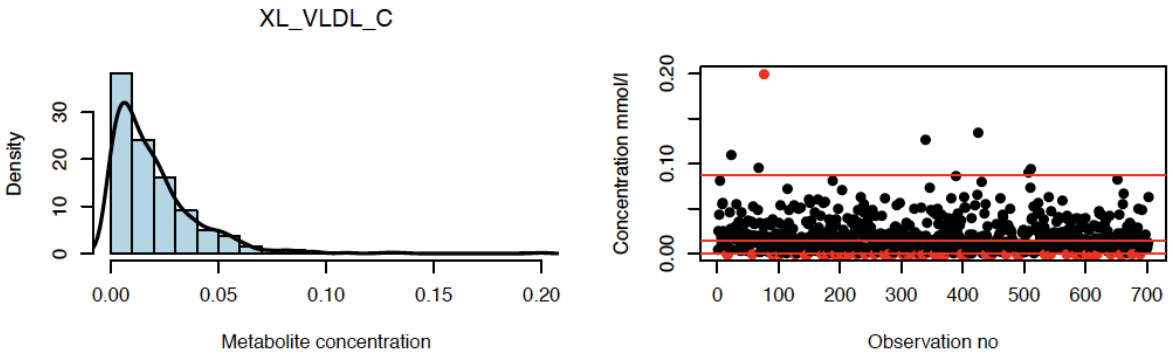


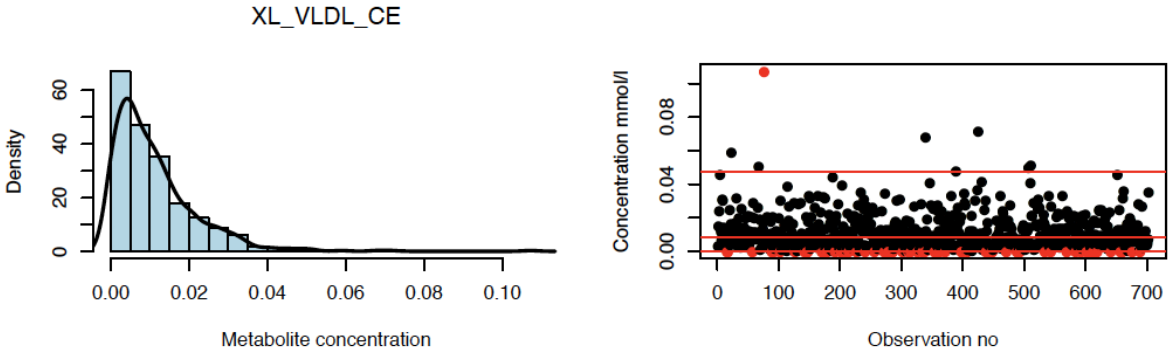
D4 continued.



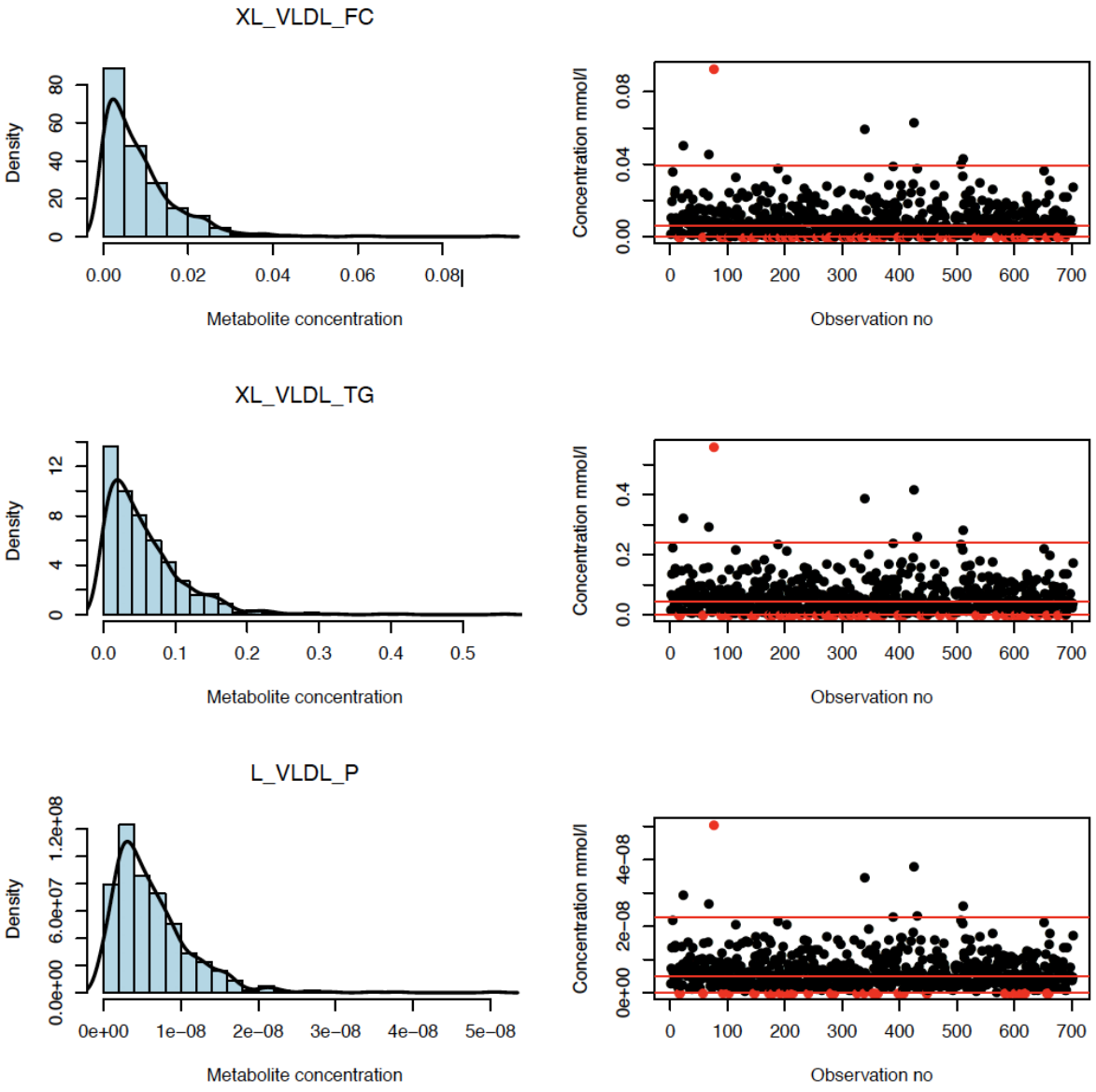
D4 continued.



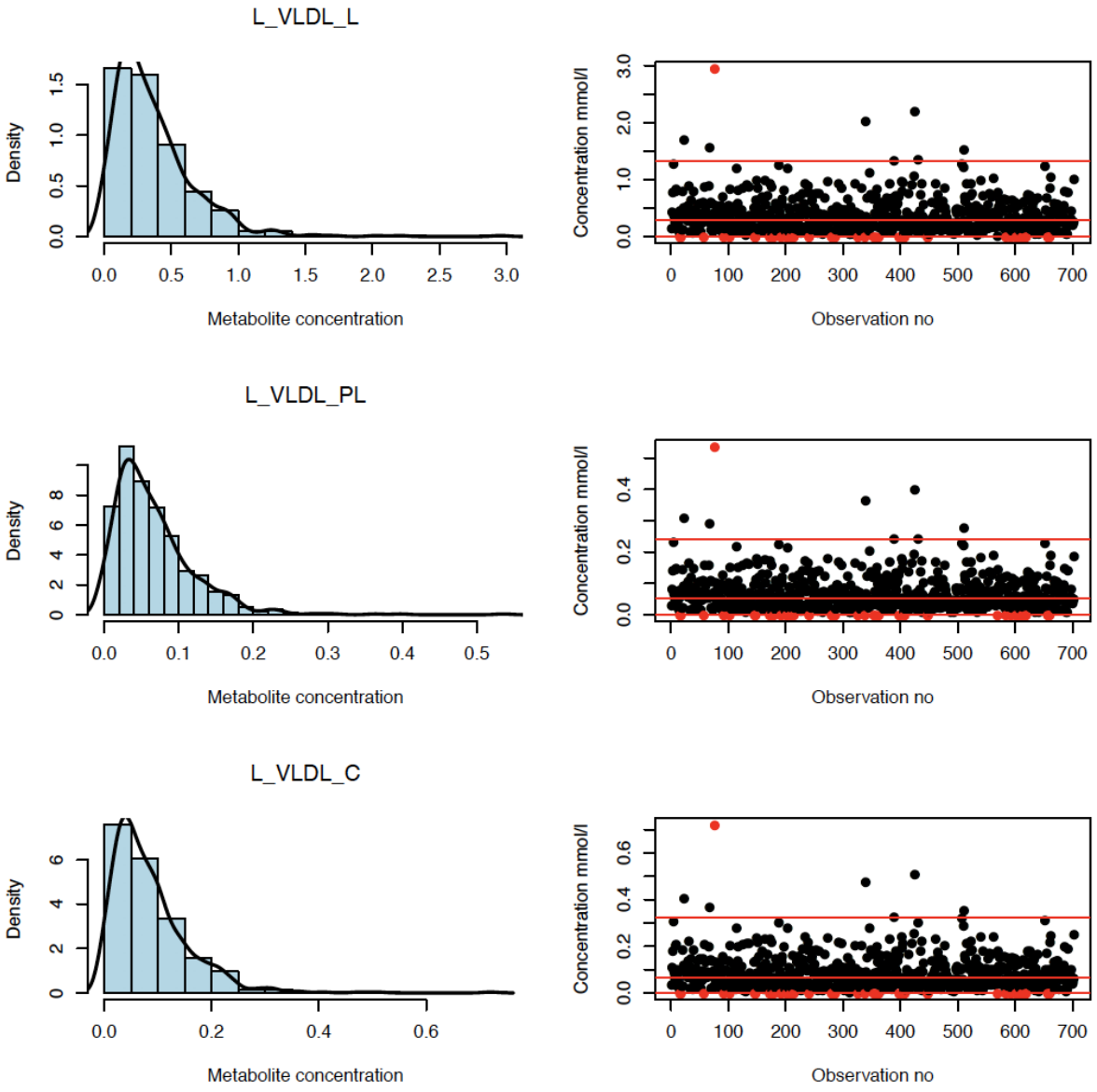




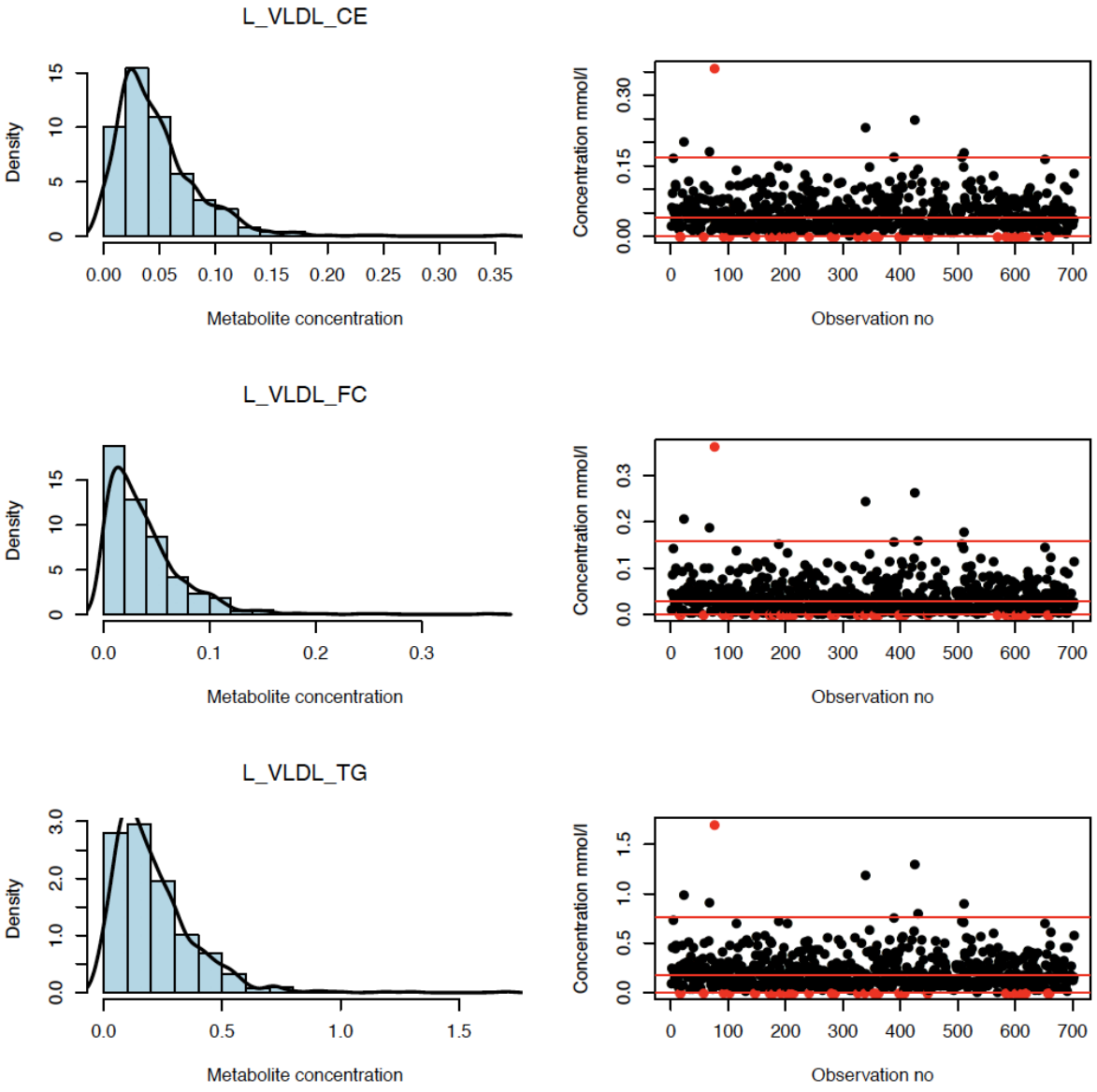
D4 continued.



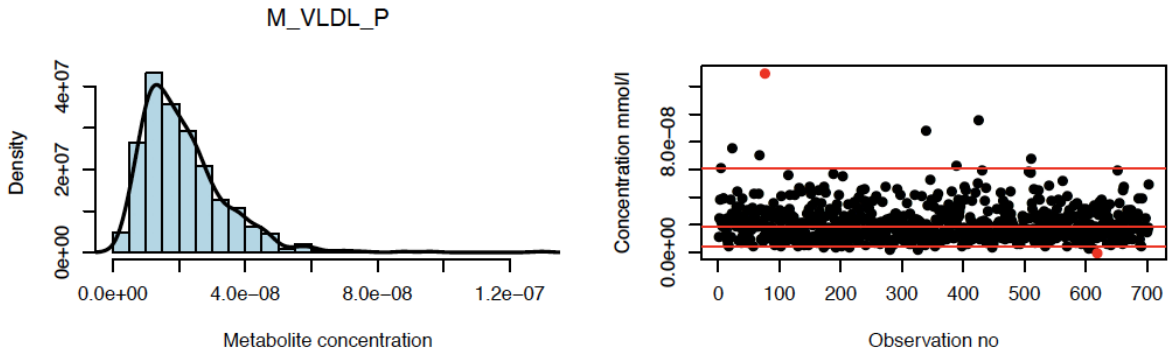
D4 continued.

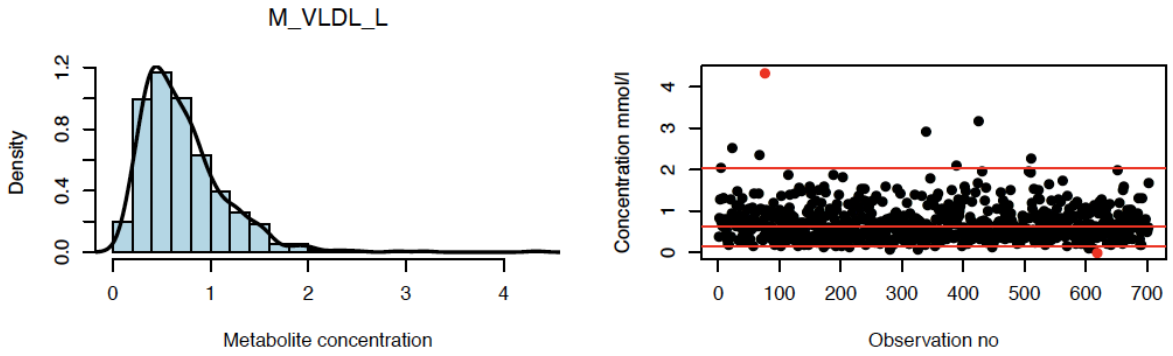


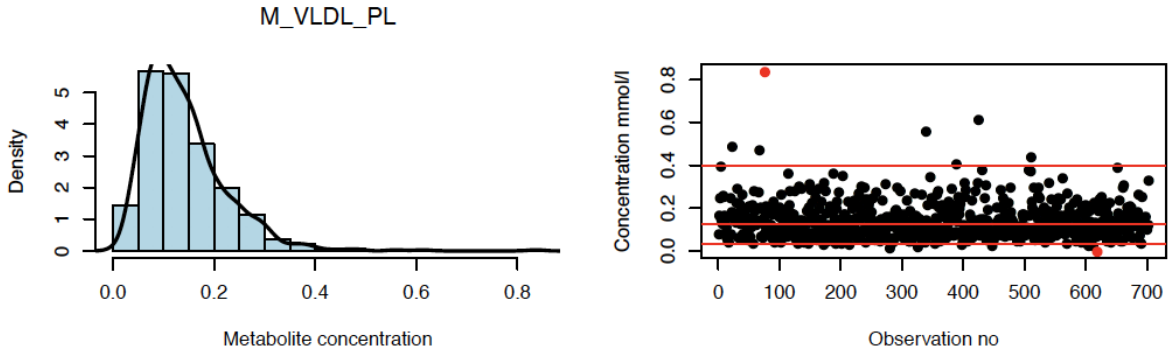
D4 continued.



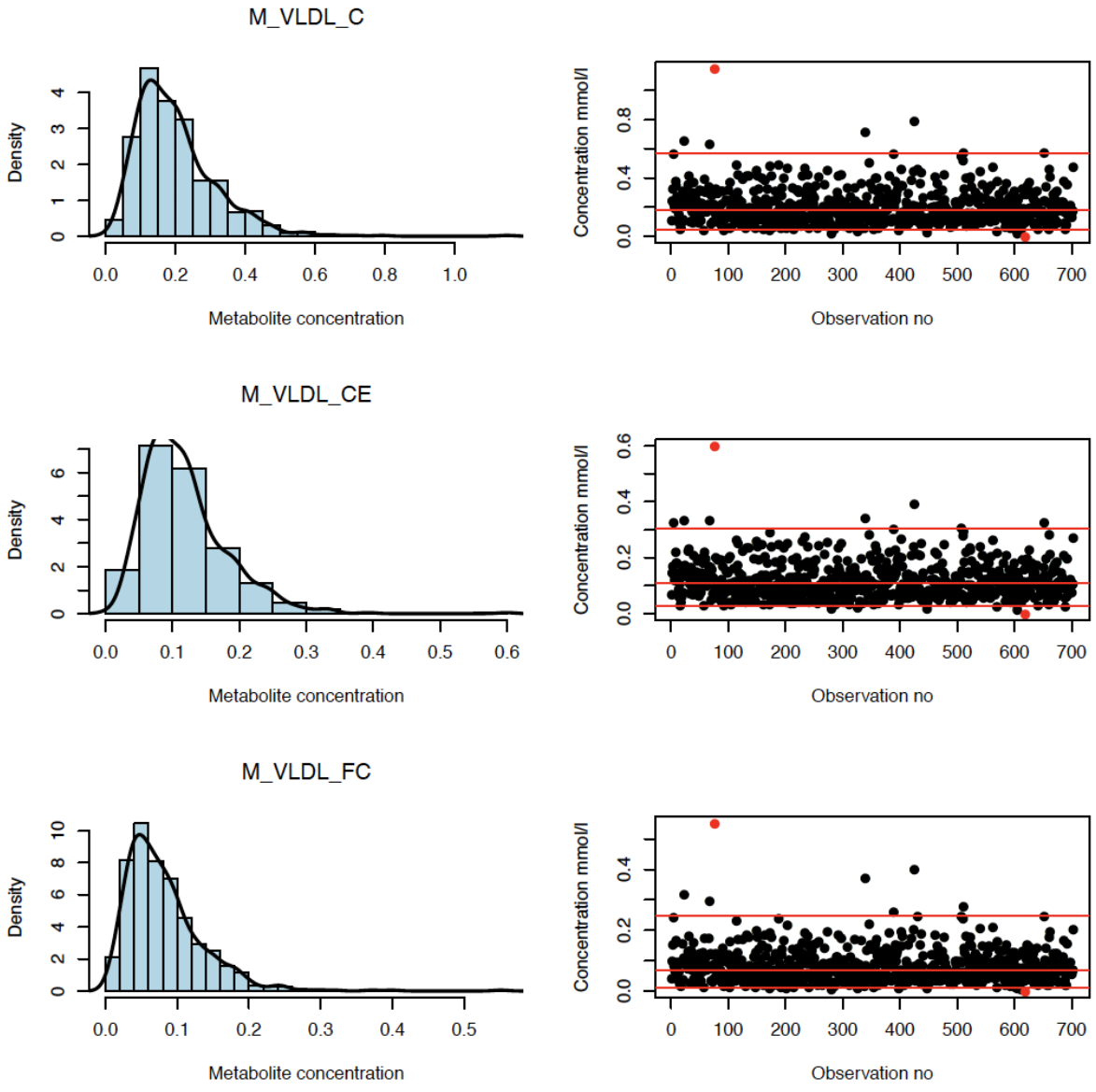
D4 continued.



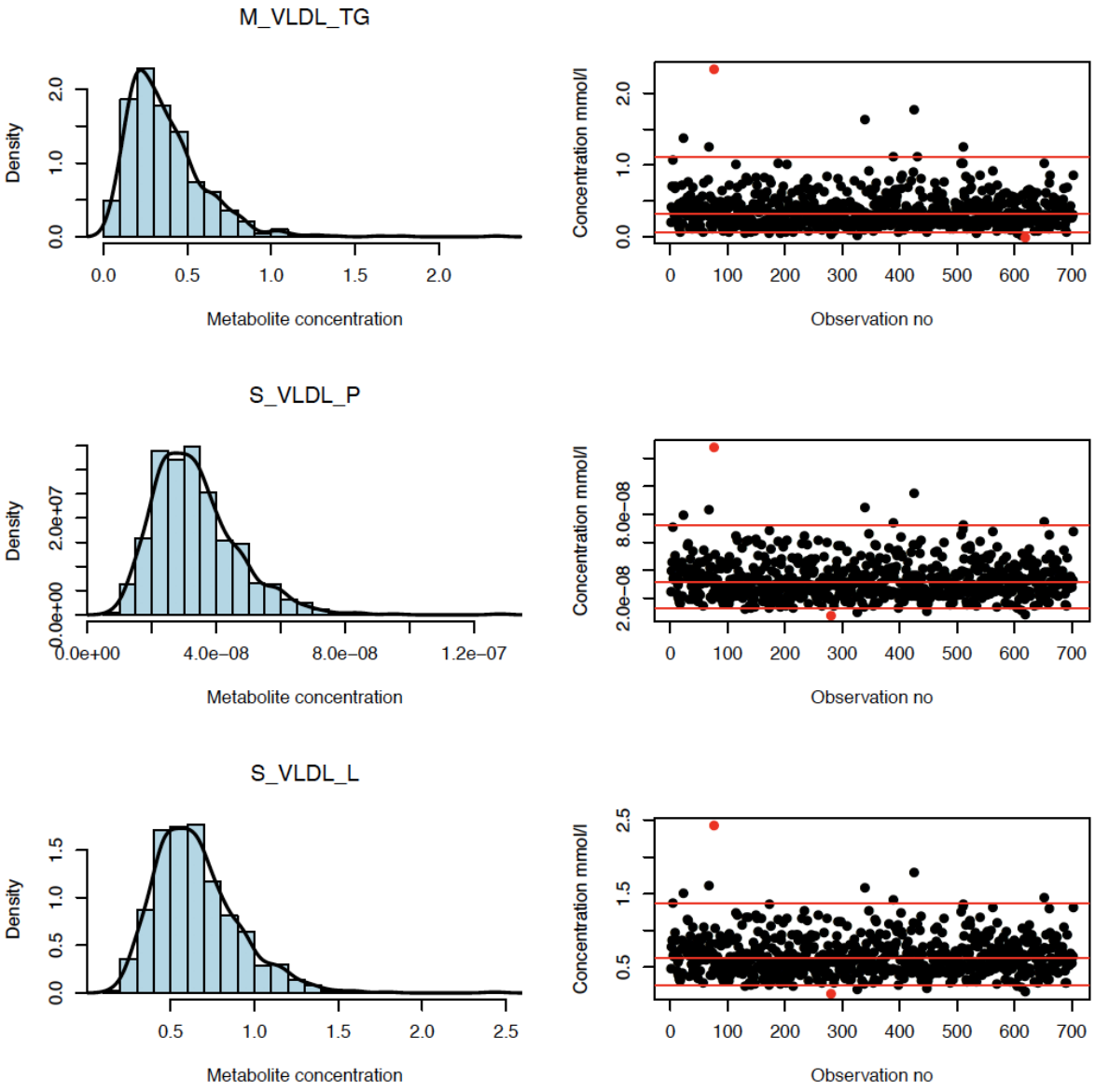




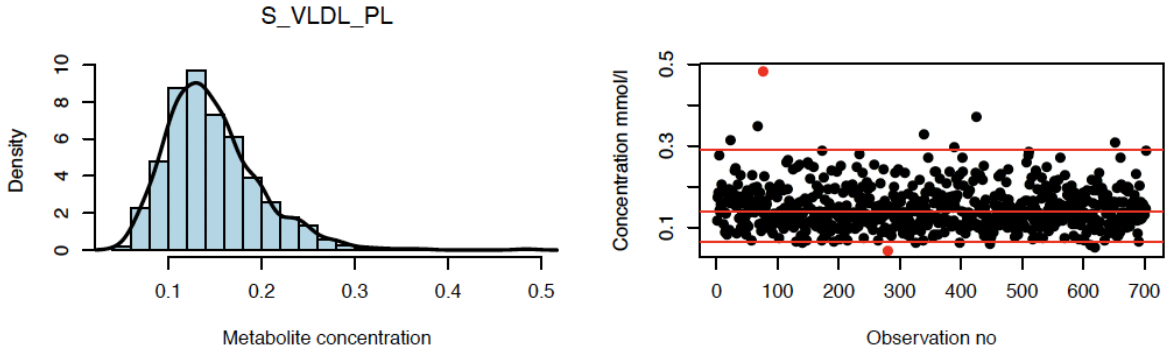
D4 continued.

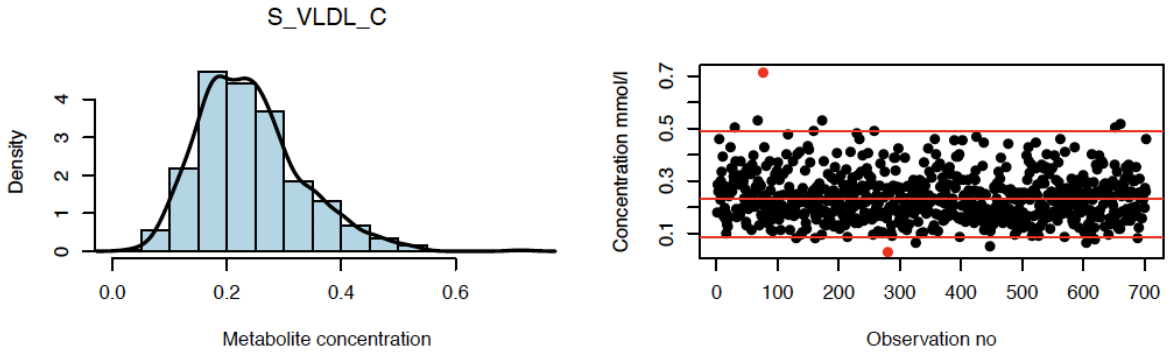


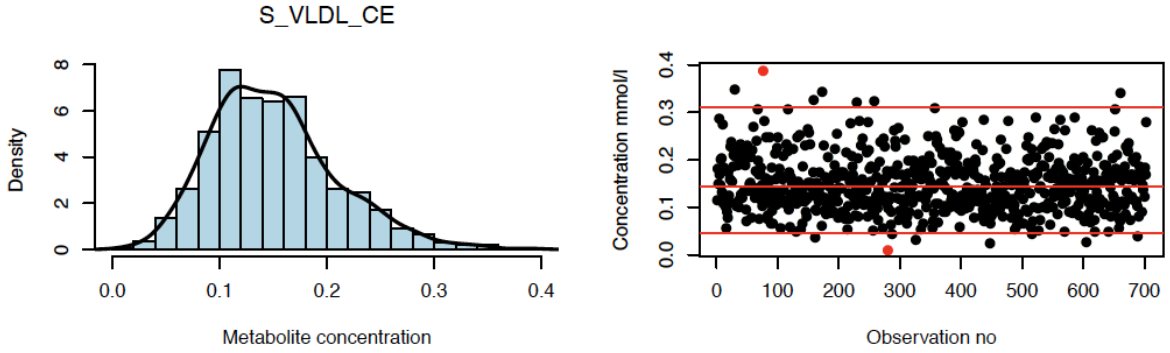
D4 continued.



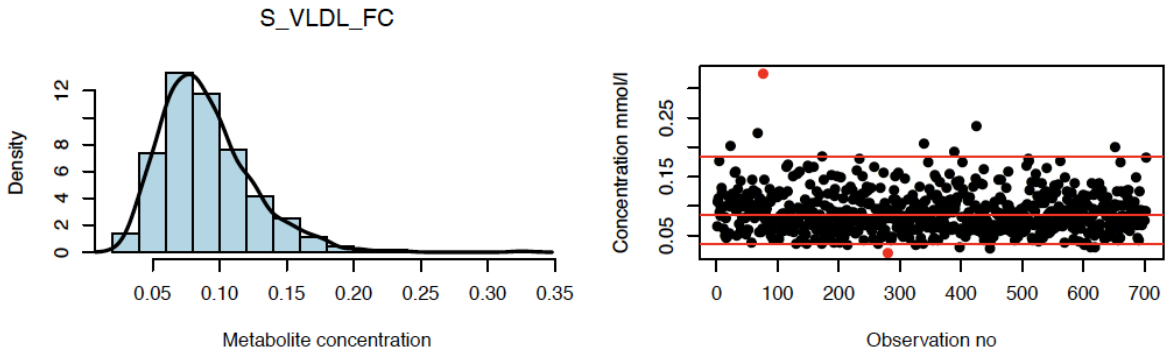
D4 continued.

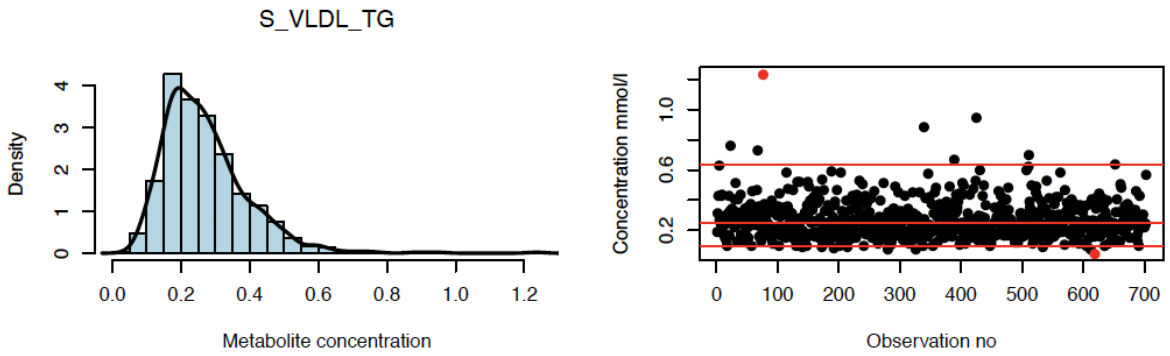


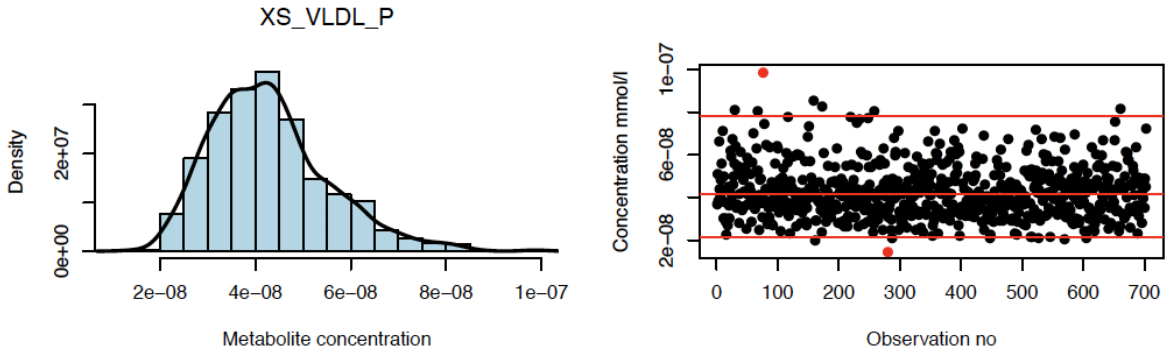




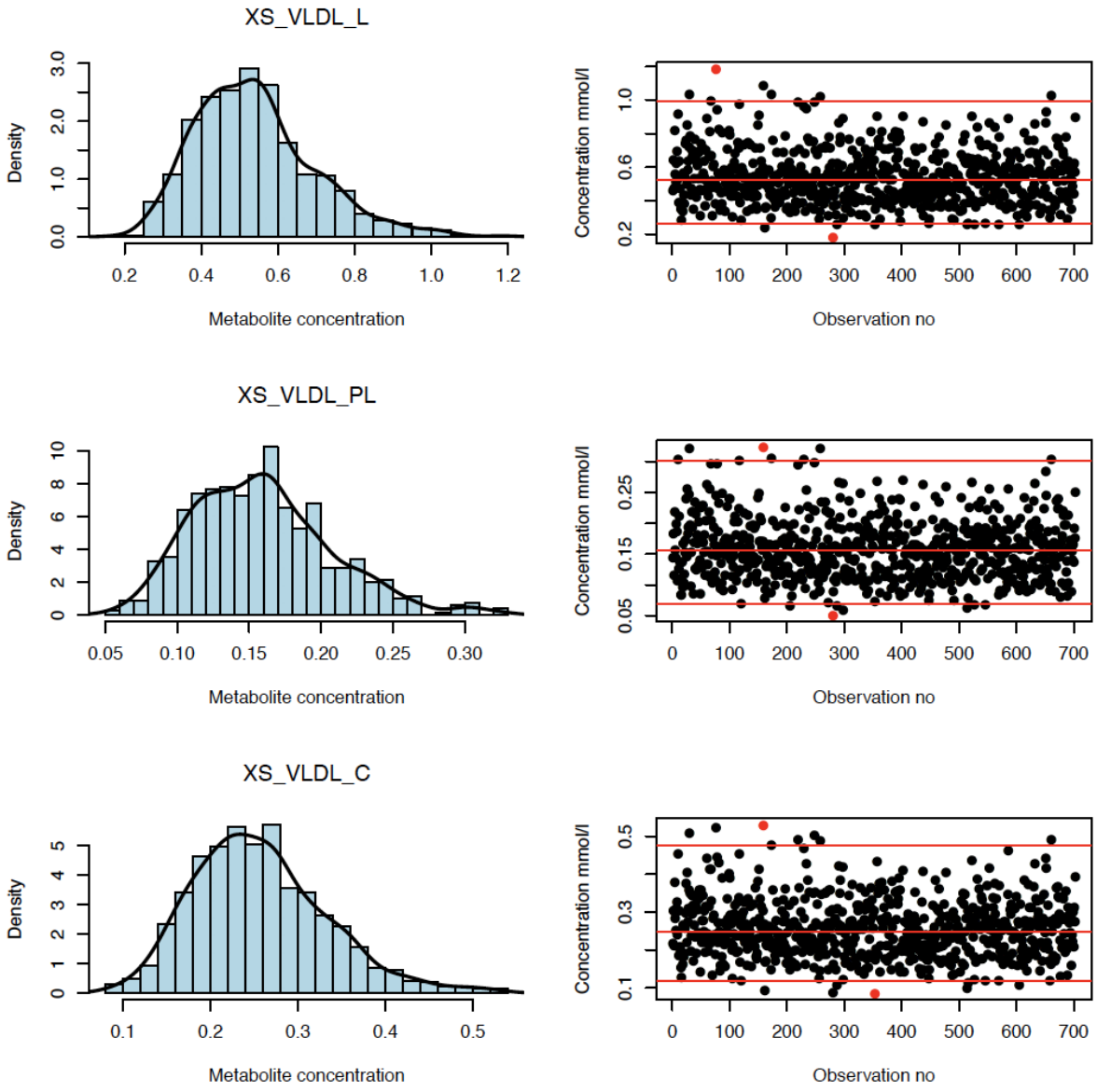
D4 continued.



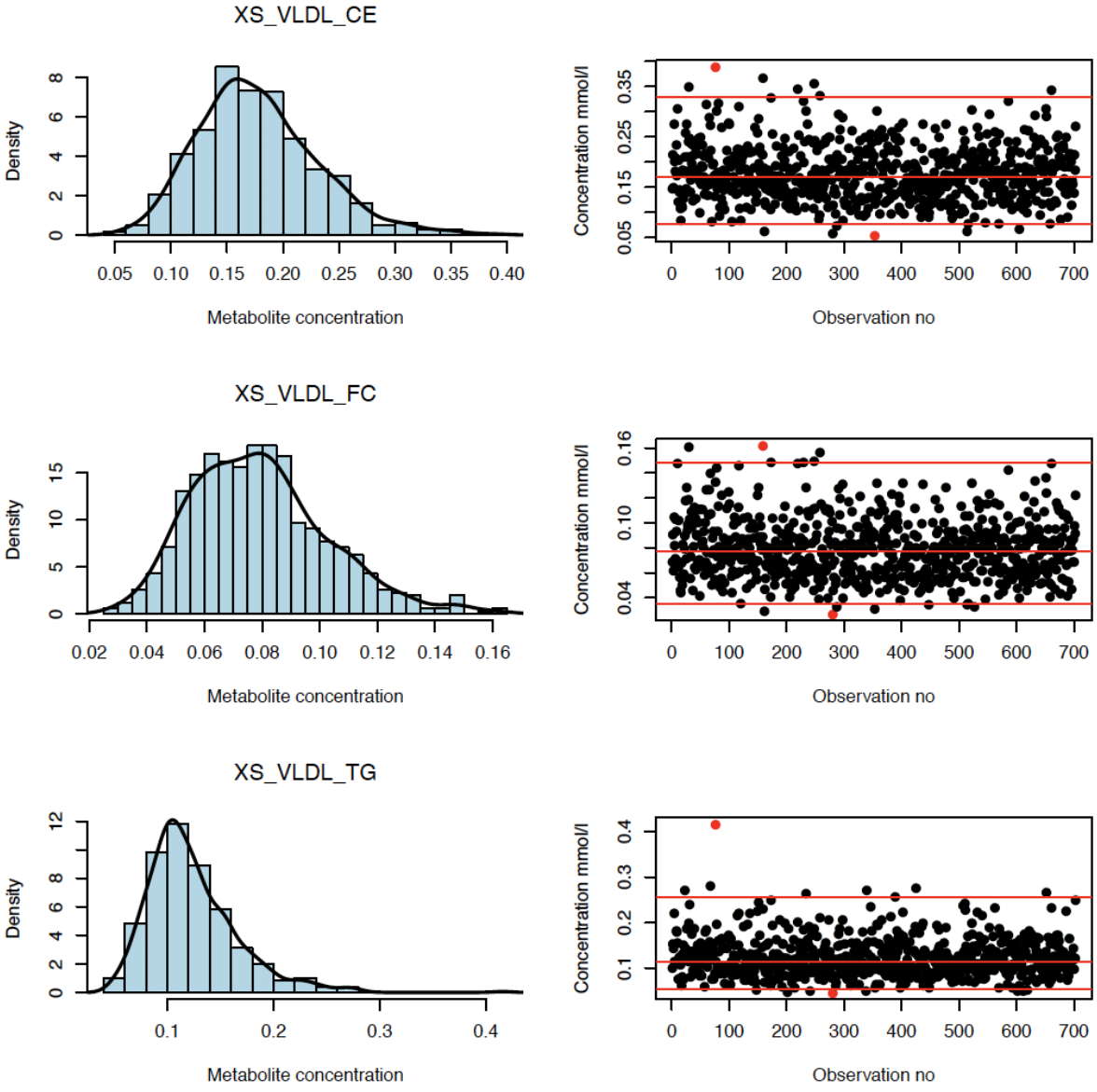




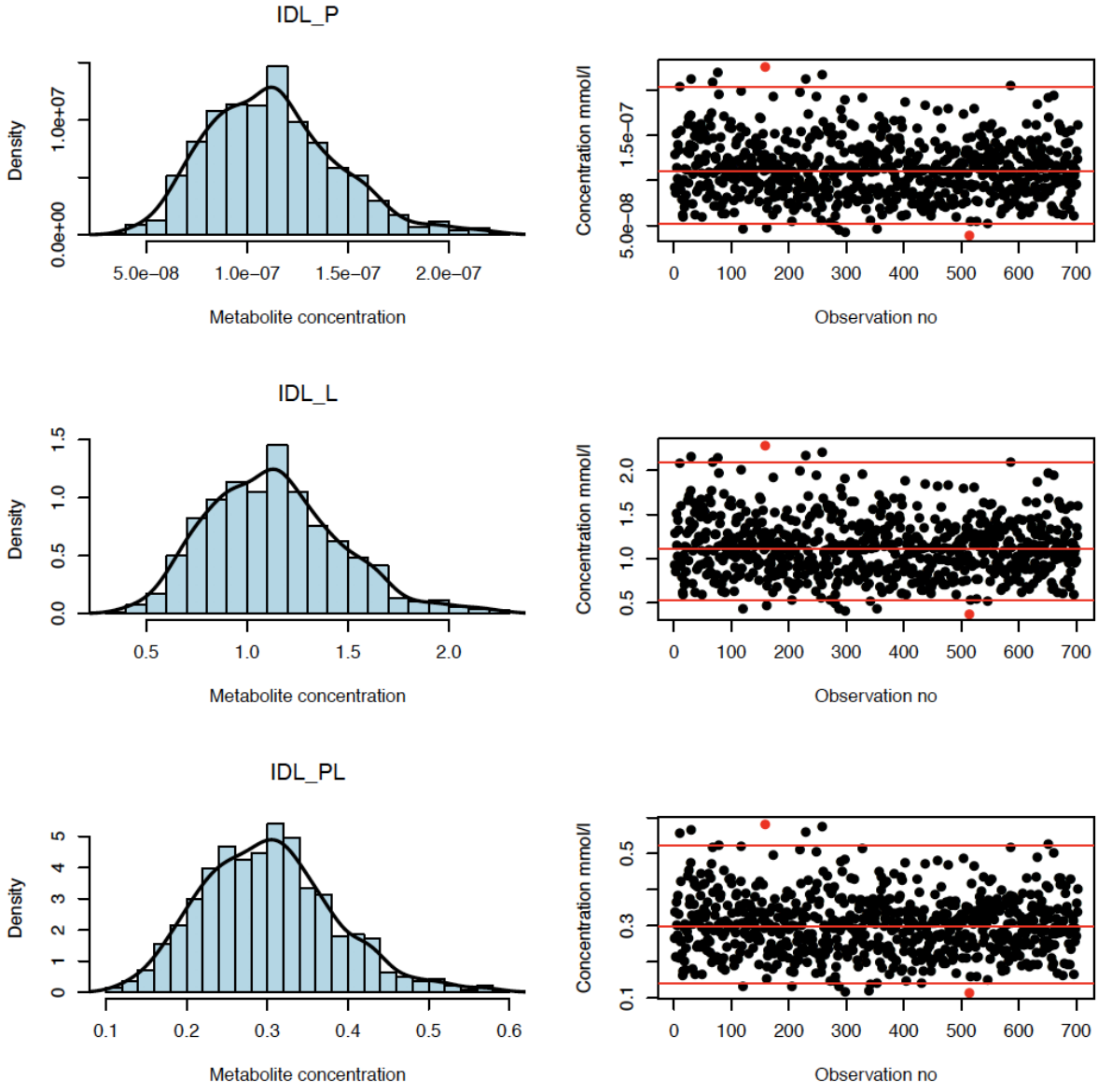
D4 continued.



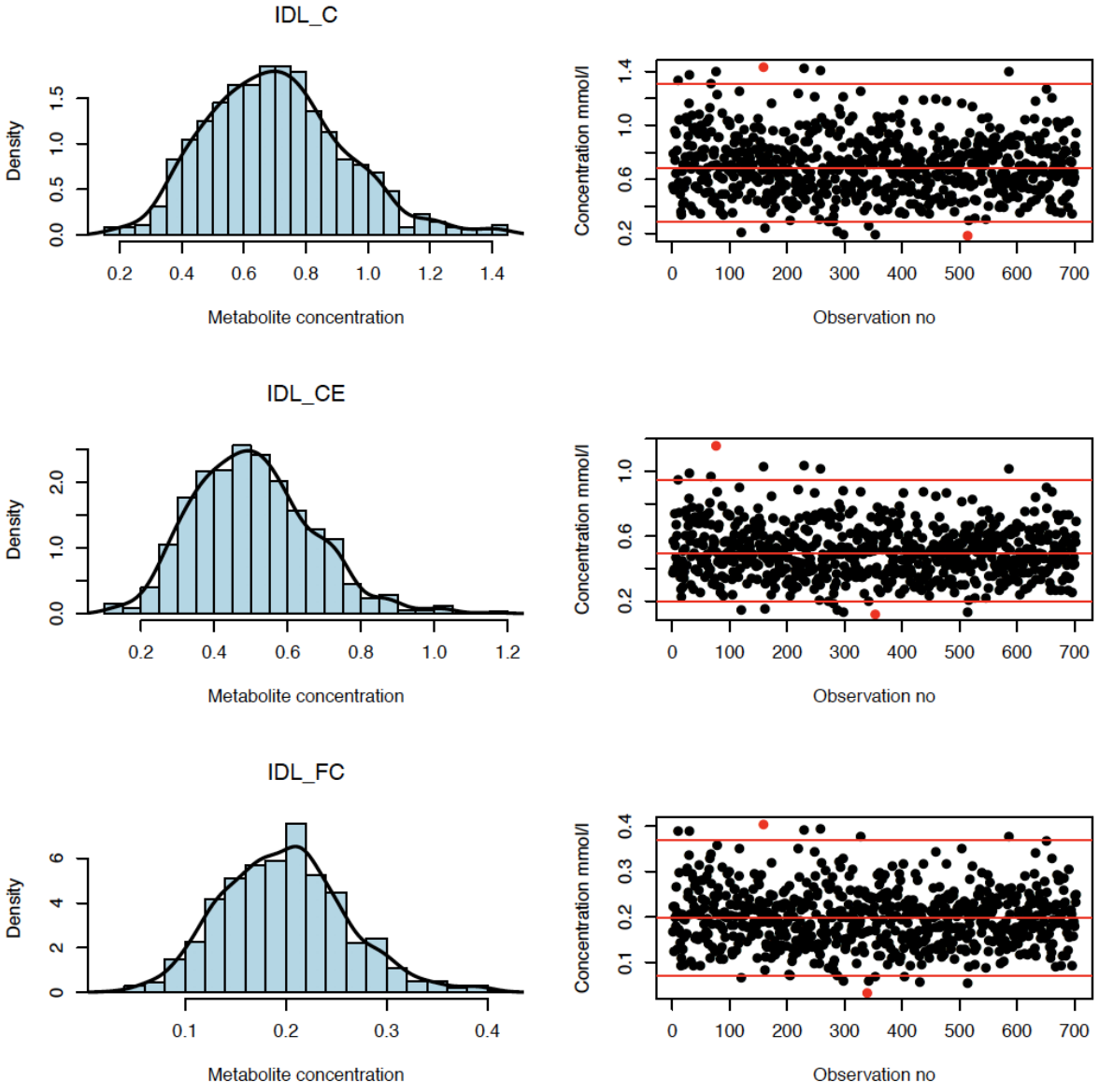
D4 continued.



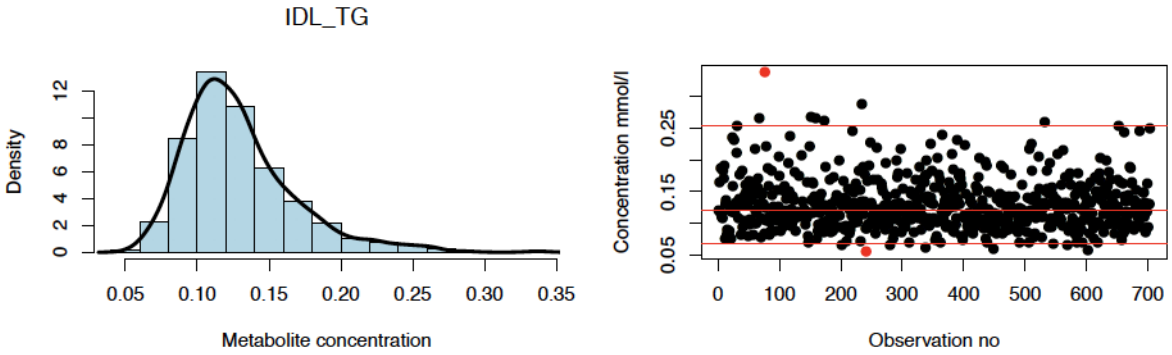
D4 continued.

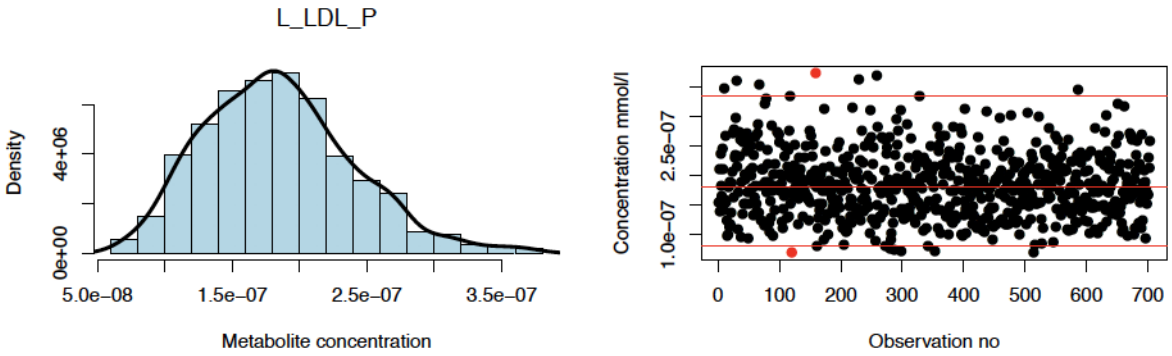


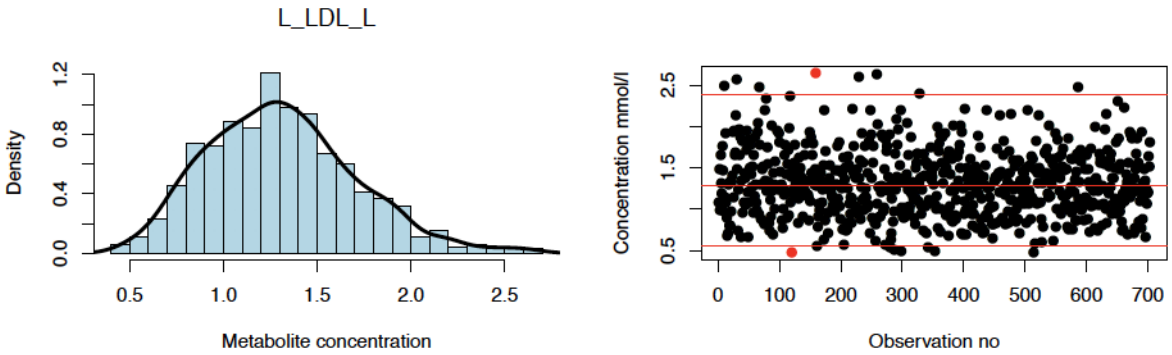
D4 continued.



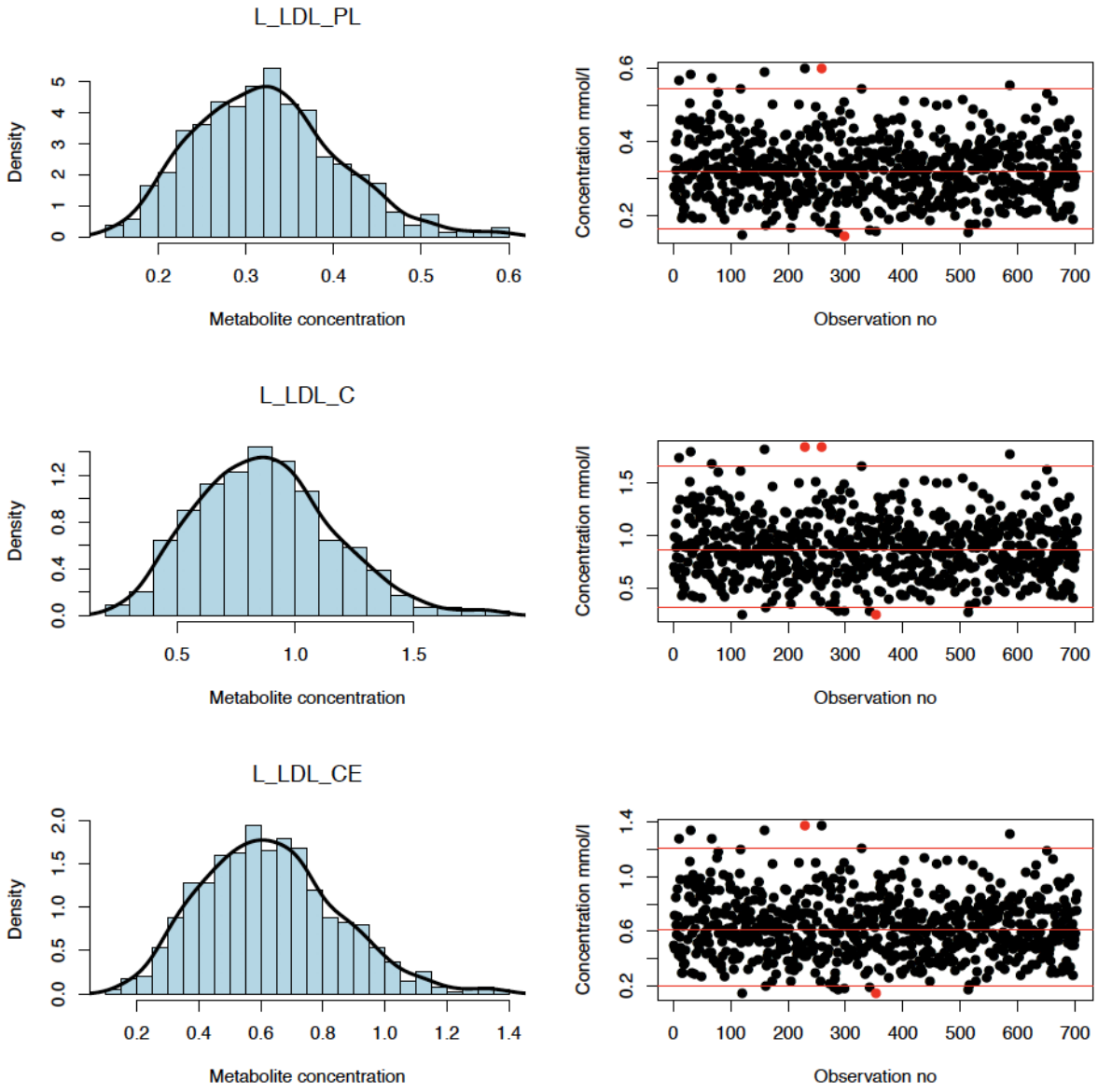
D4 continued.



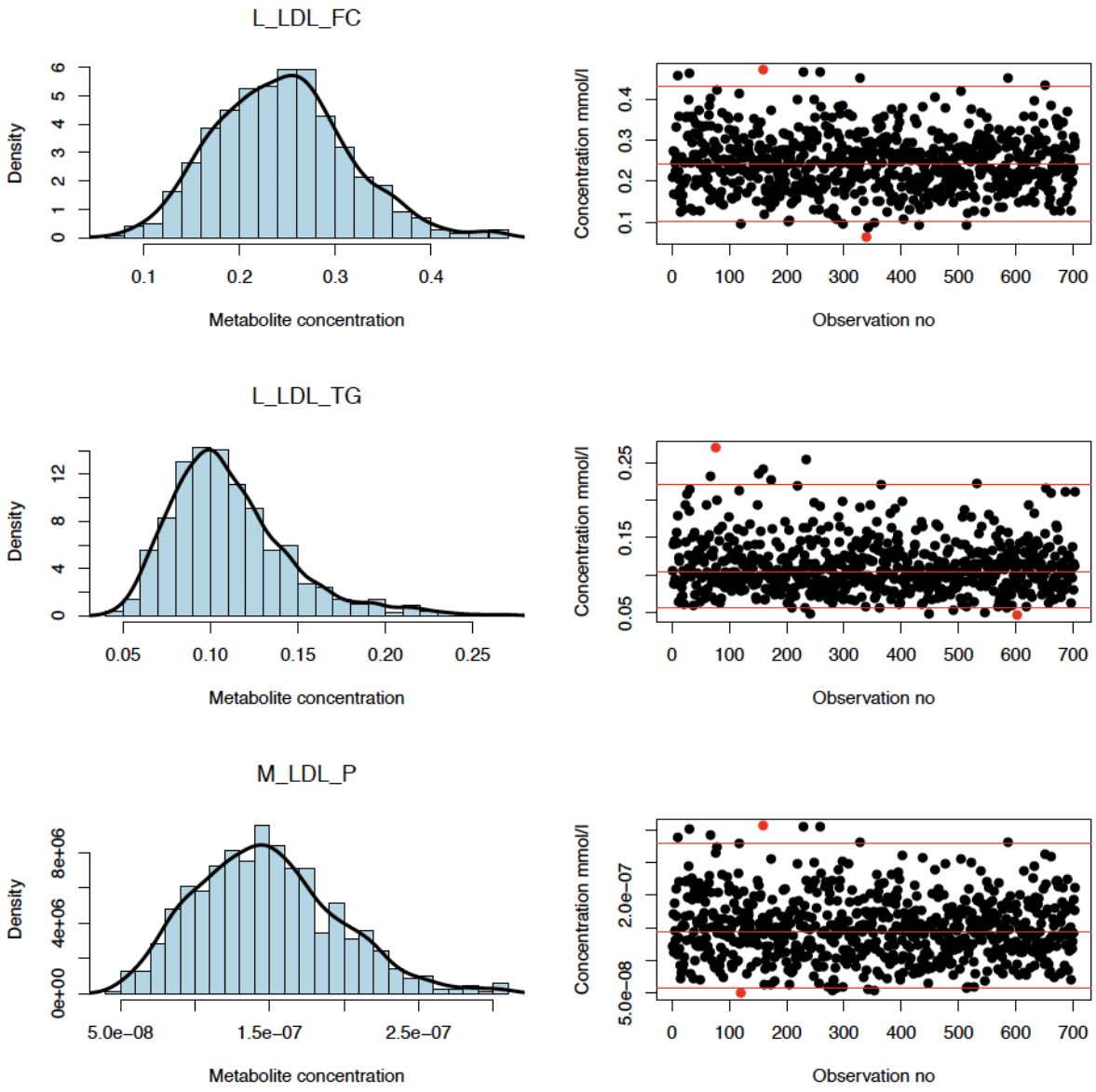




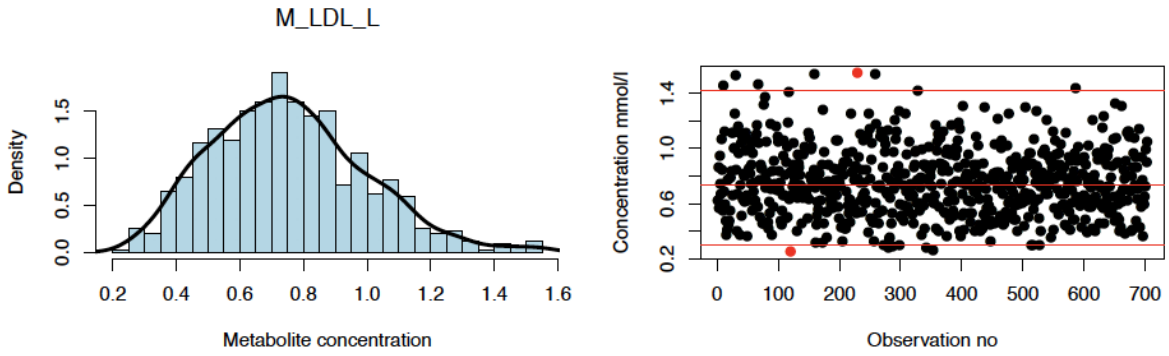
D4 continued.

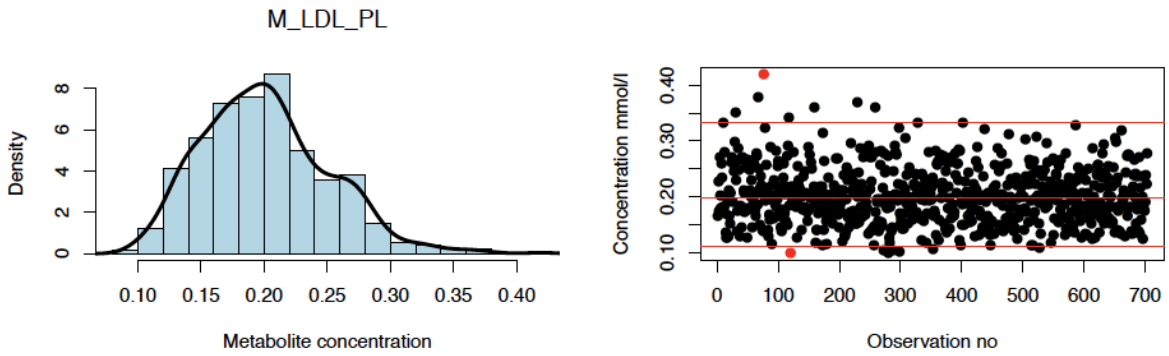


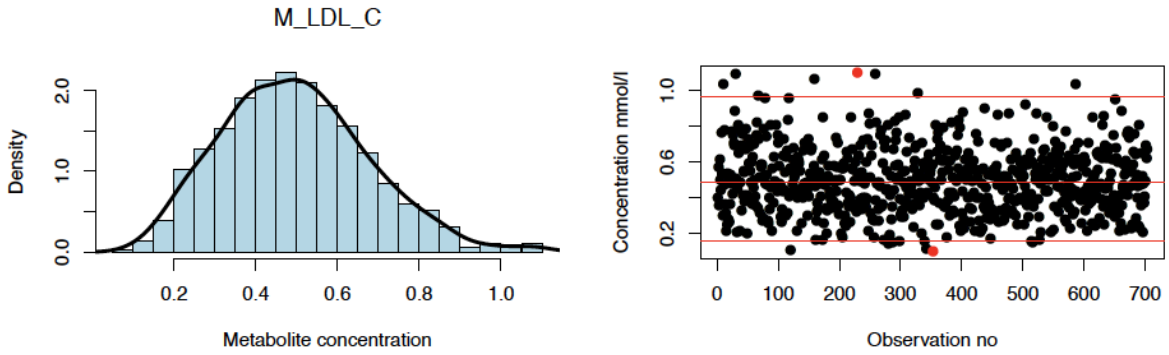
D4 continued.



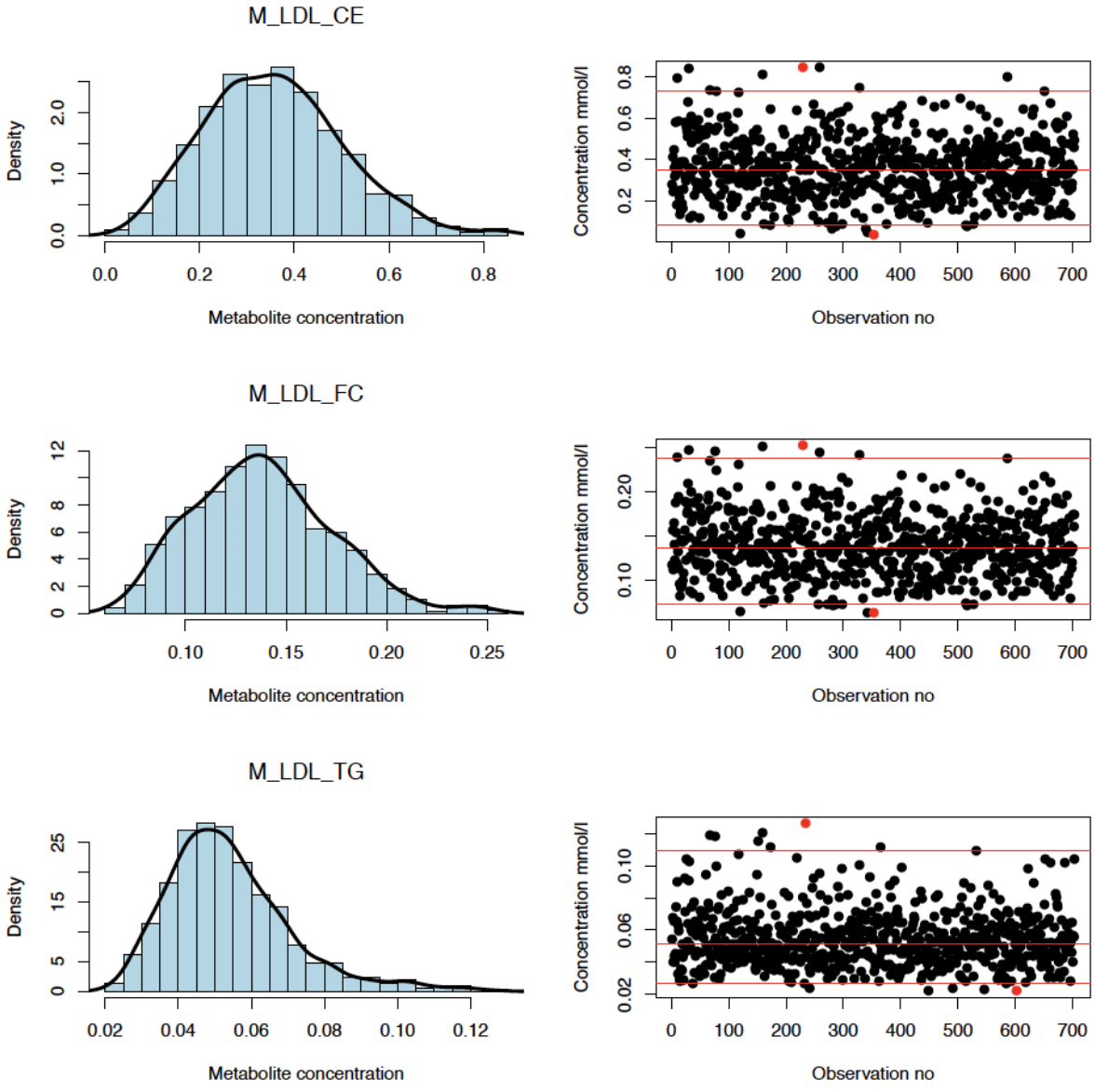
D4 continued.



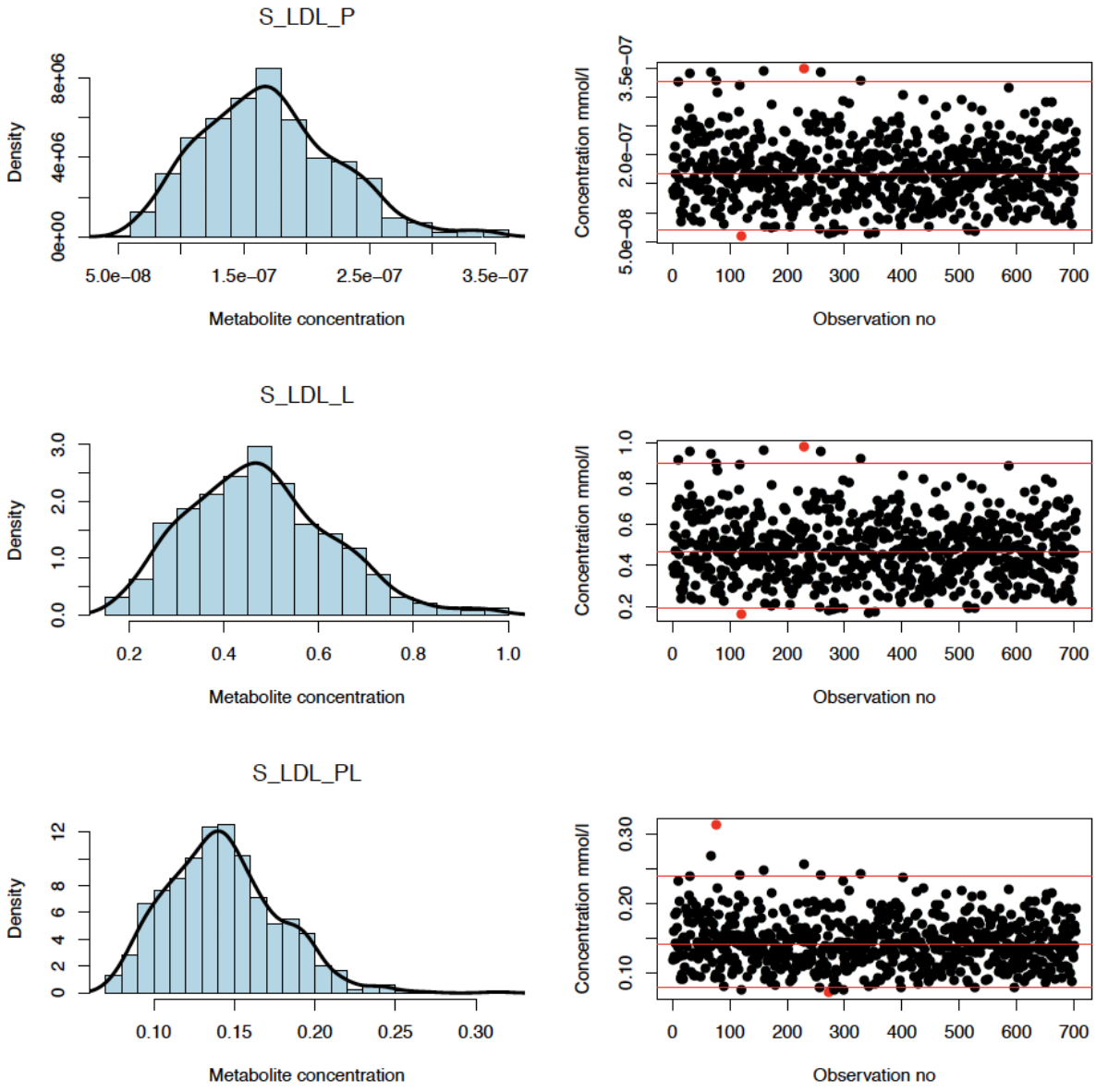




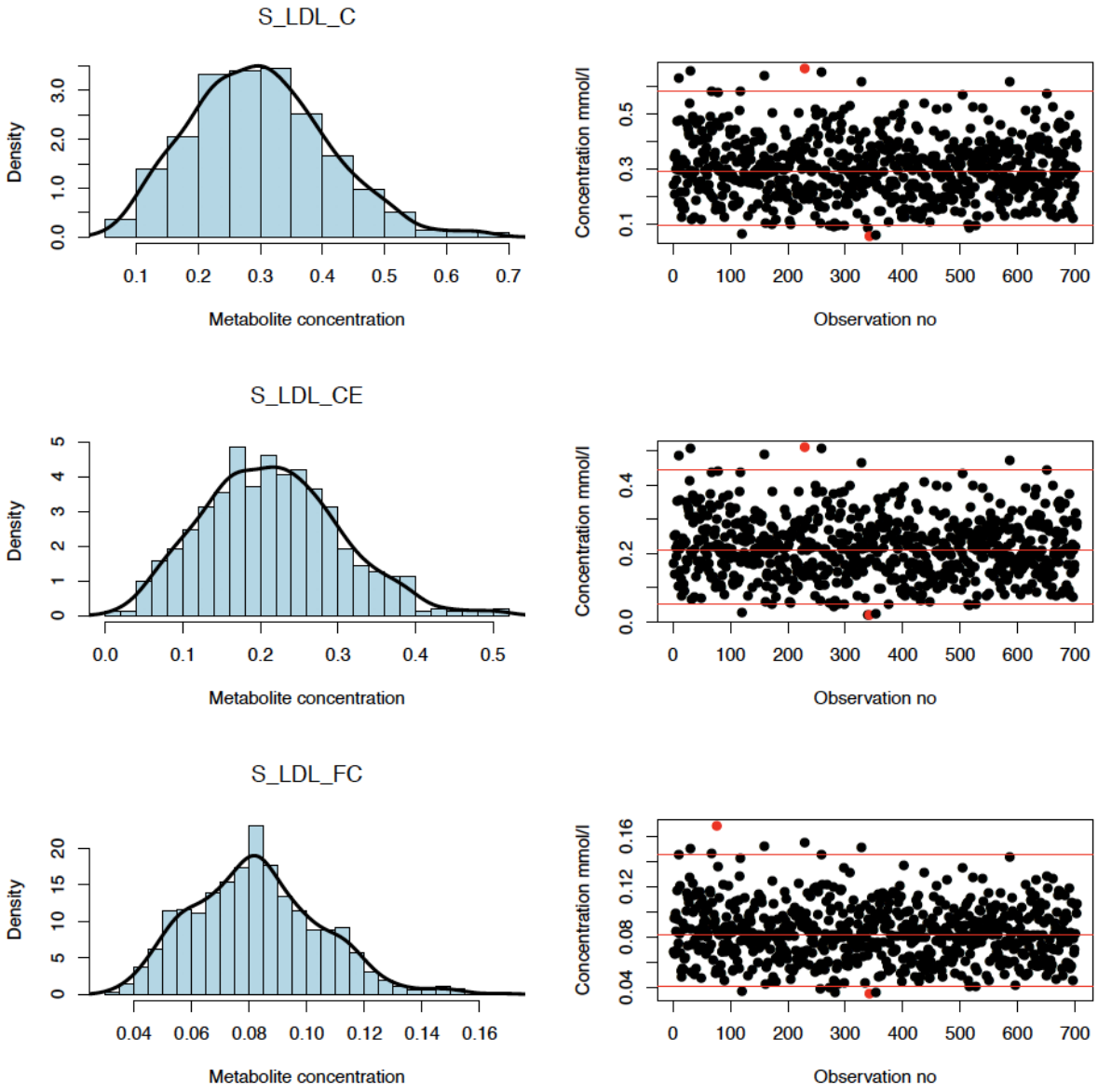
D4 continued.



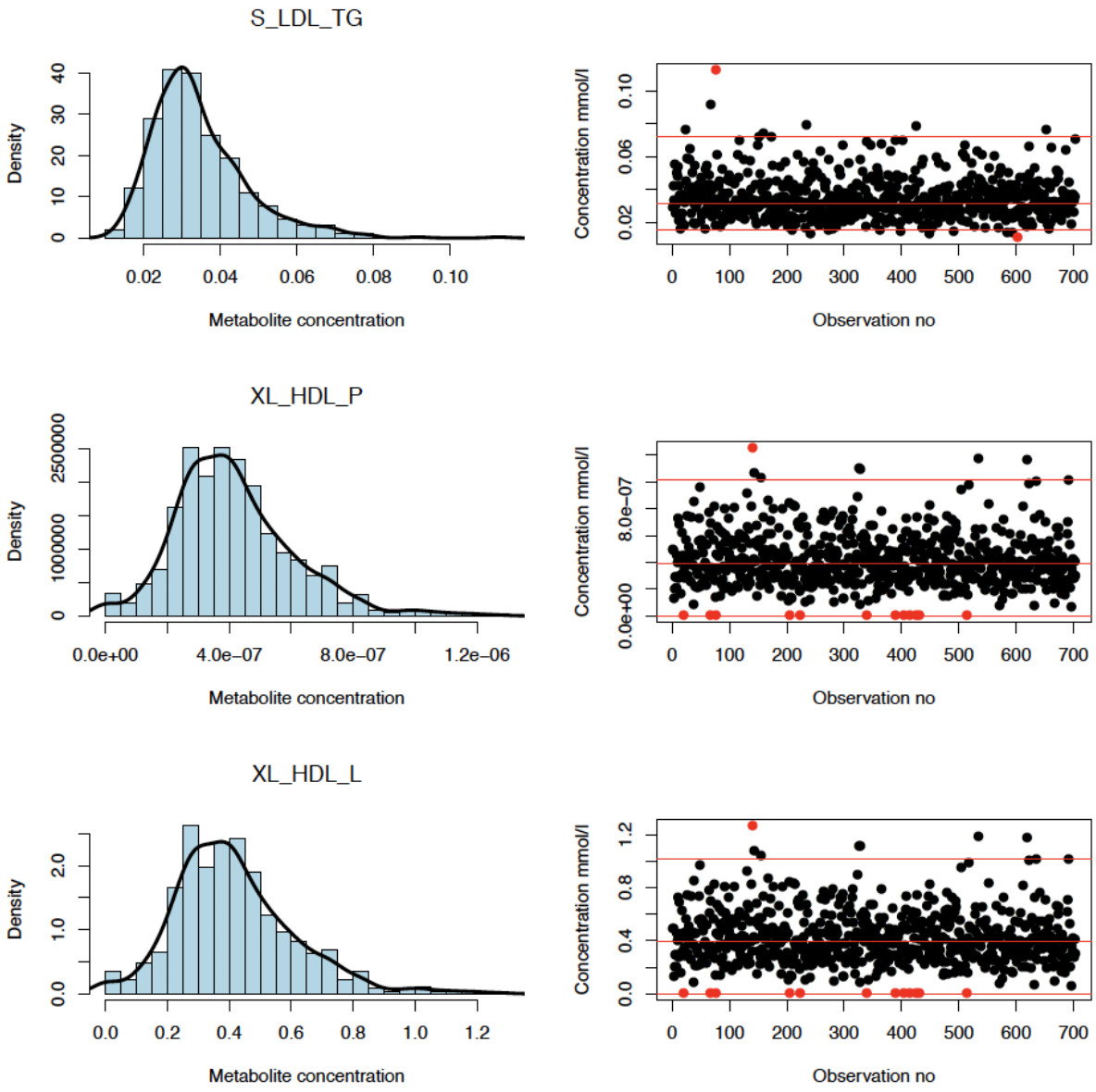
D4 continued.



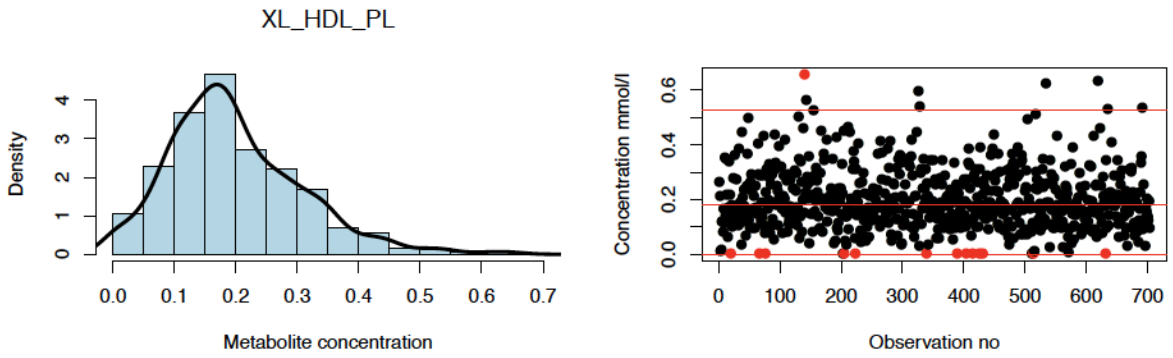
D4 continued.

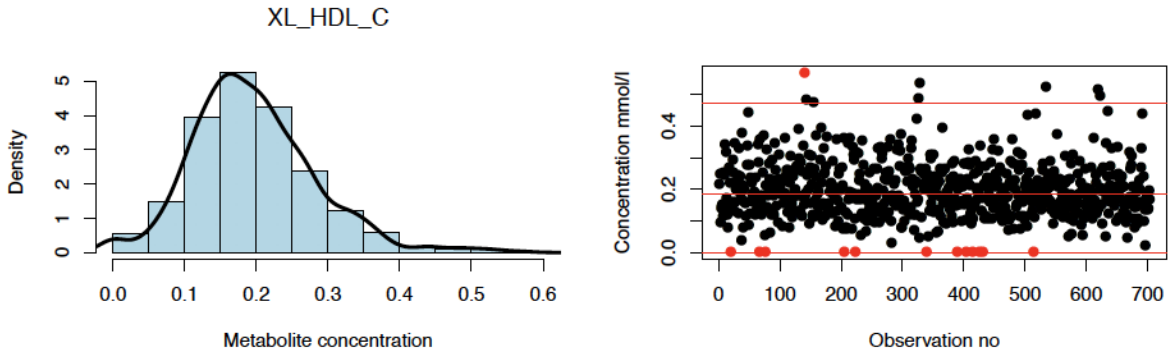


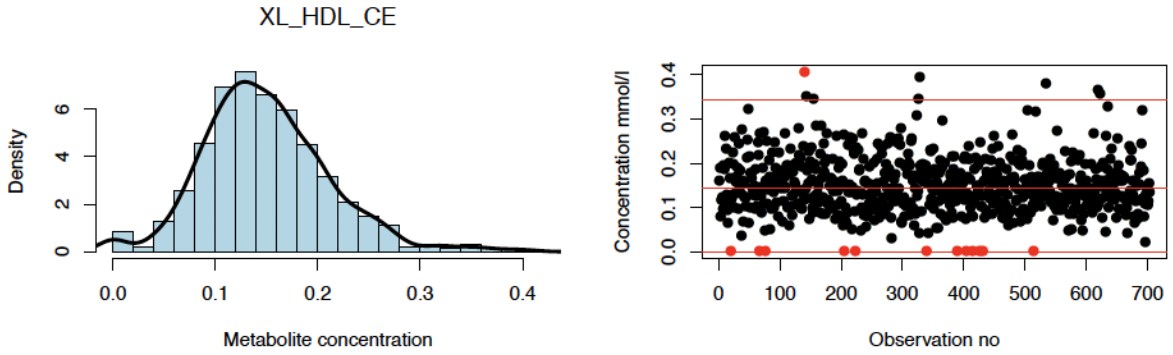
D4 continued.



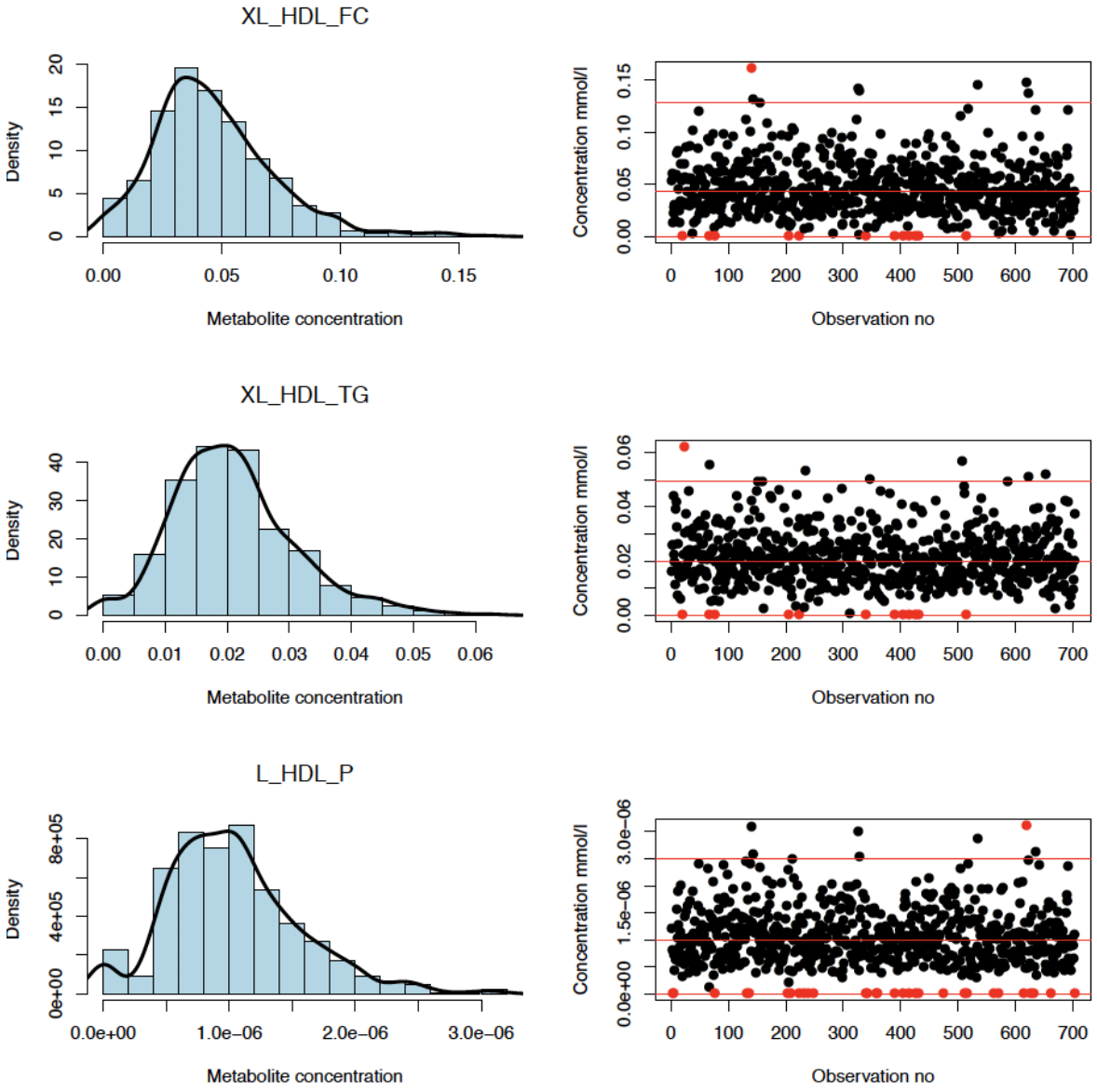
D4 continued.



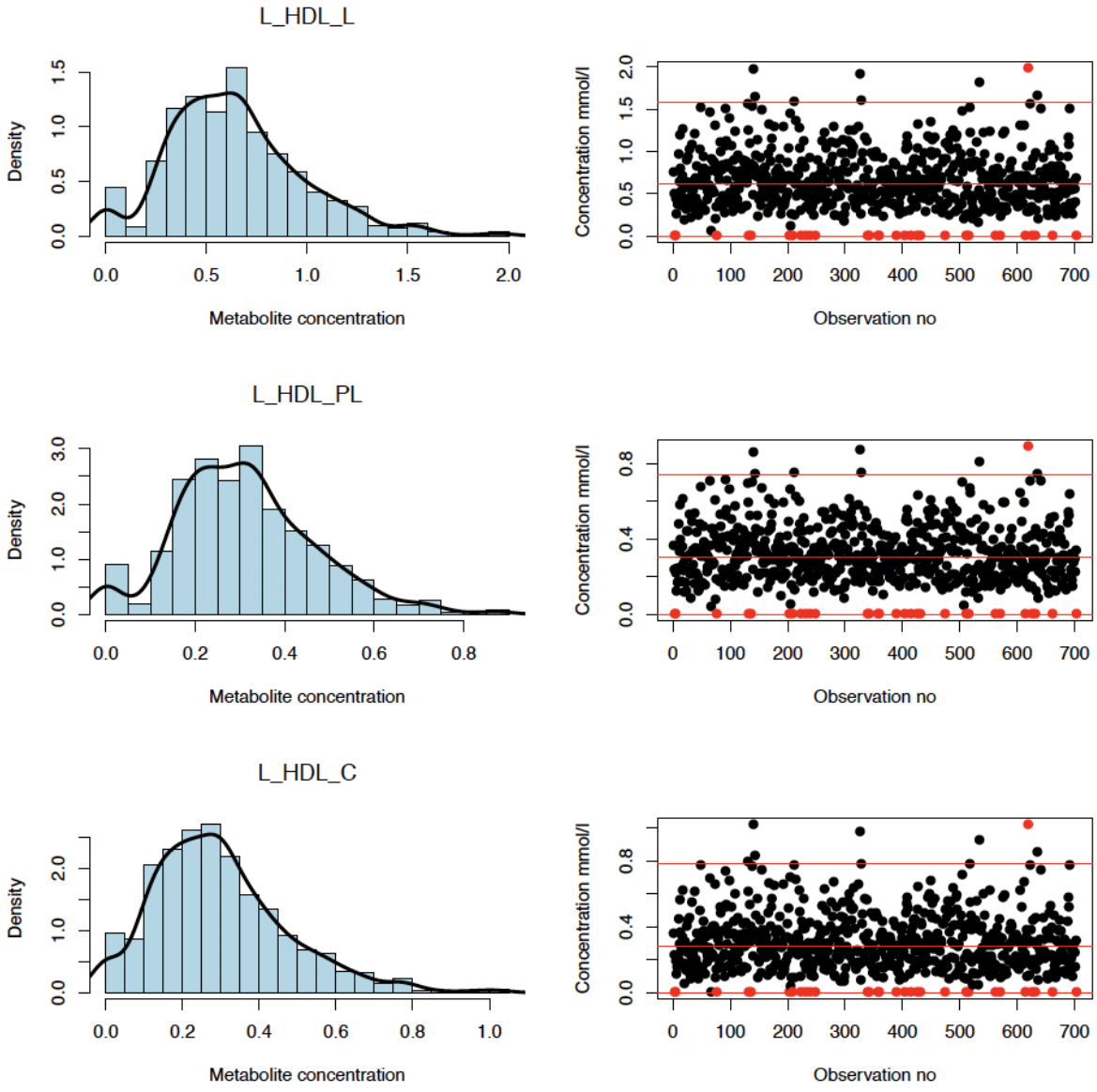




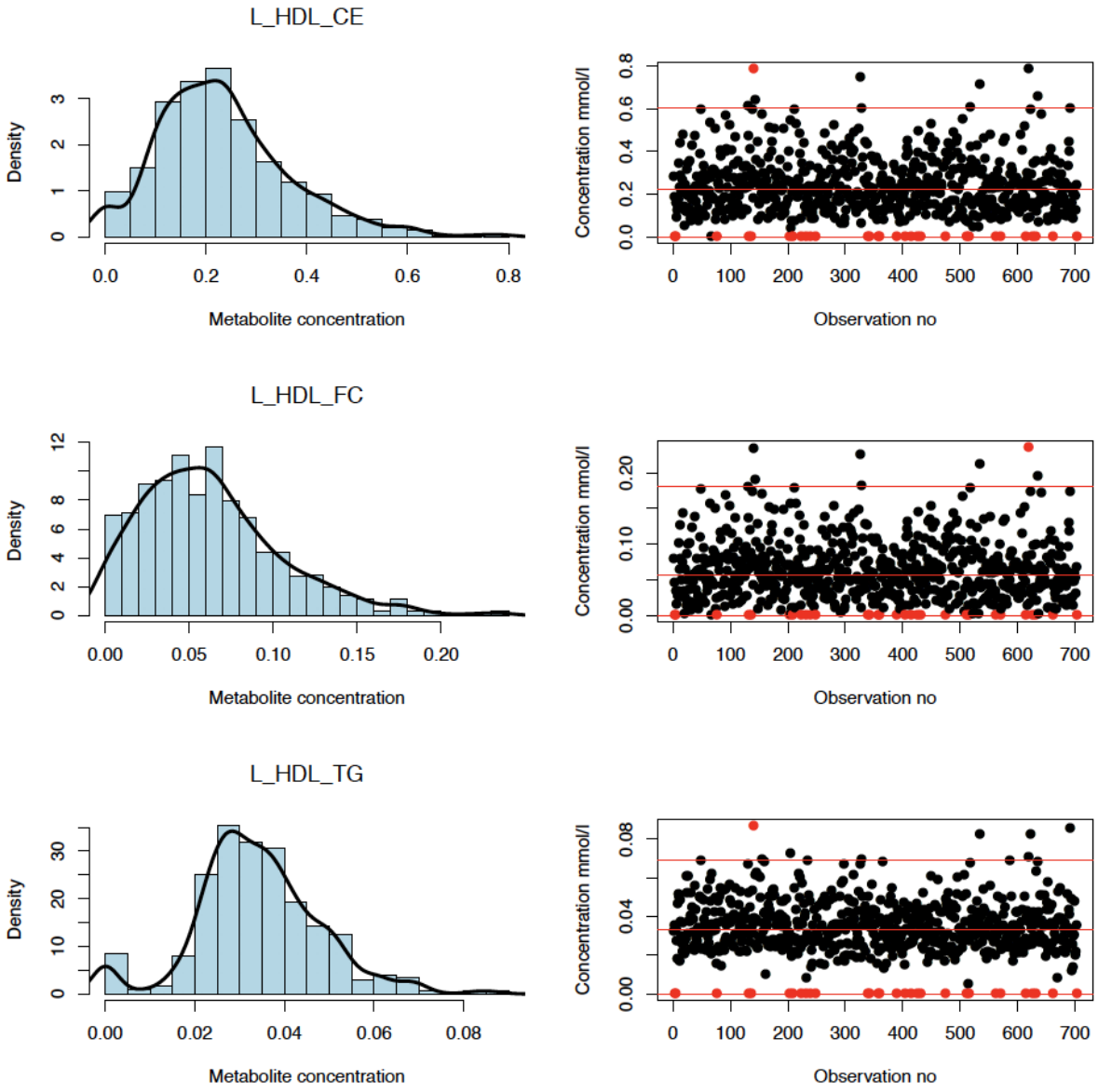
D4 continued.



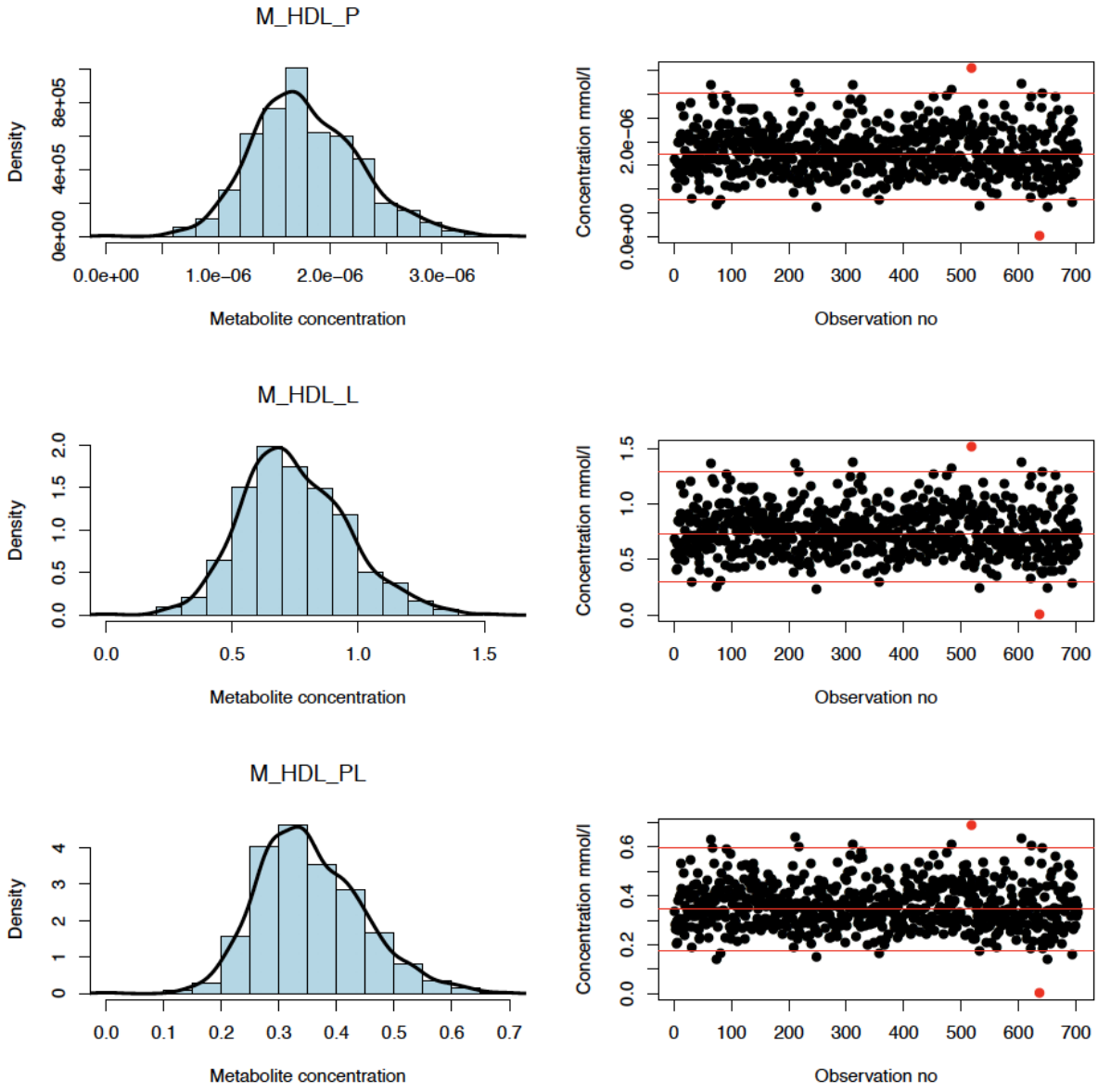
D4 continued.



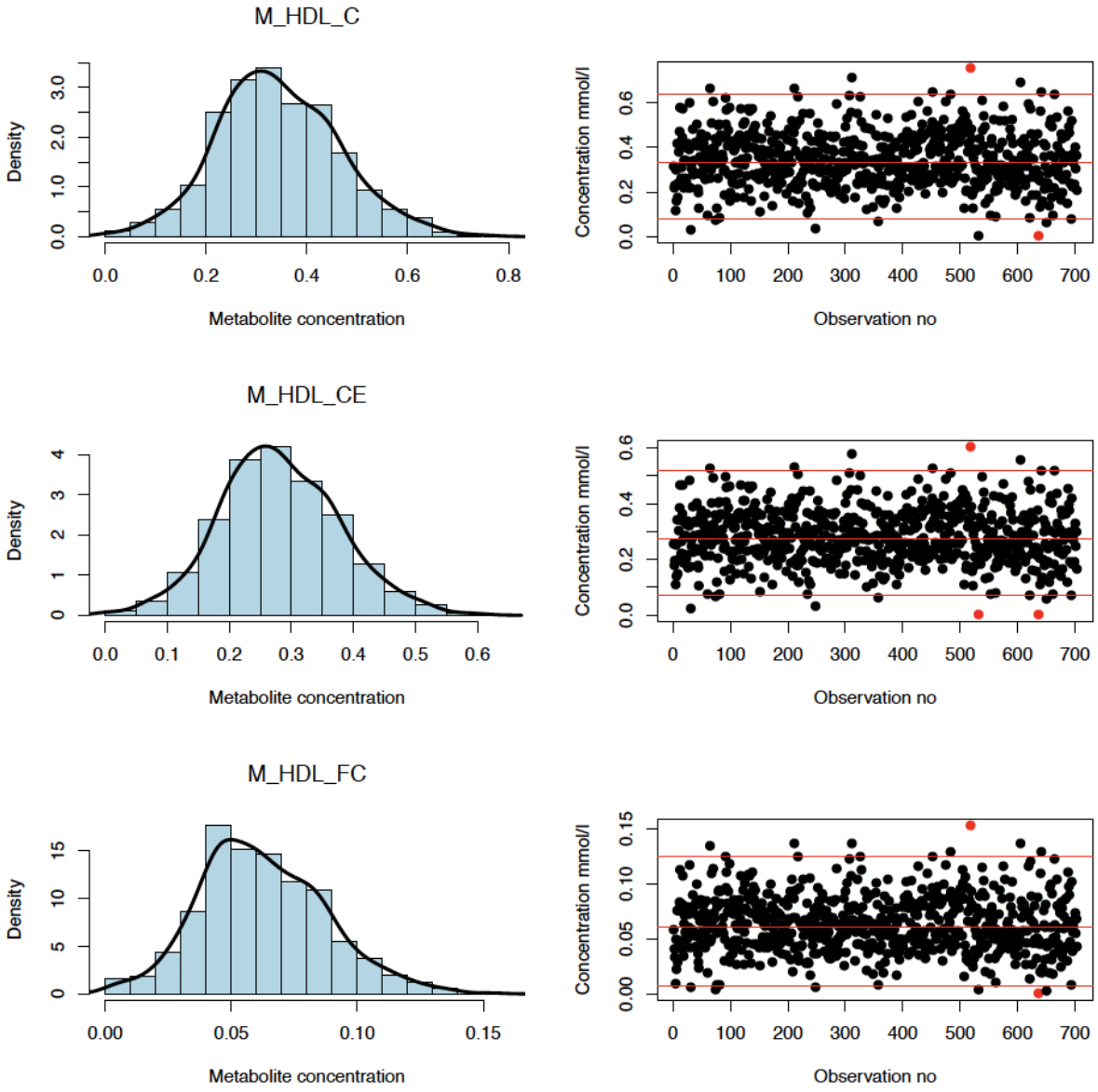
D4 continued.



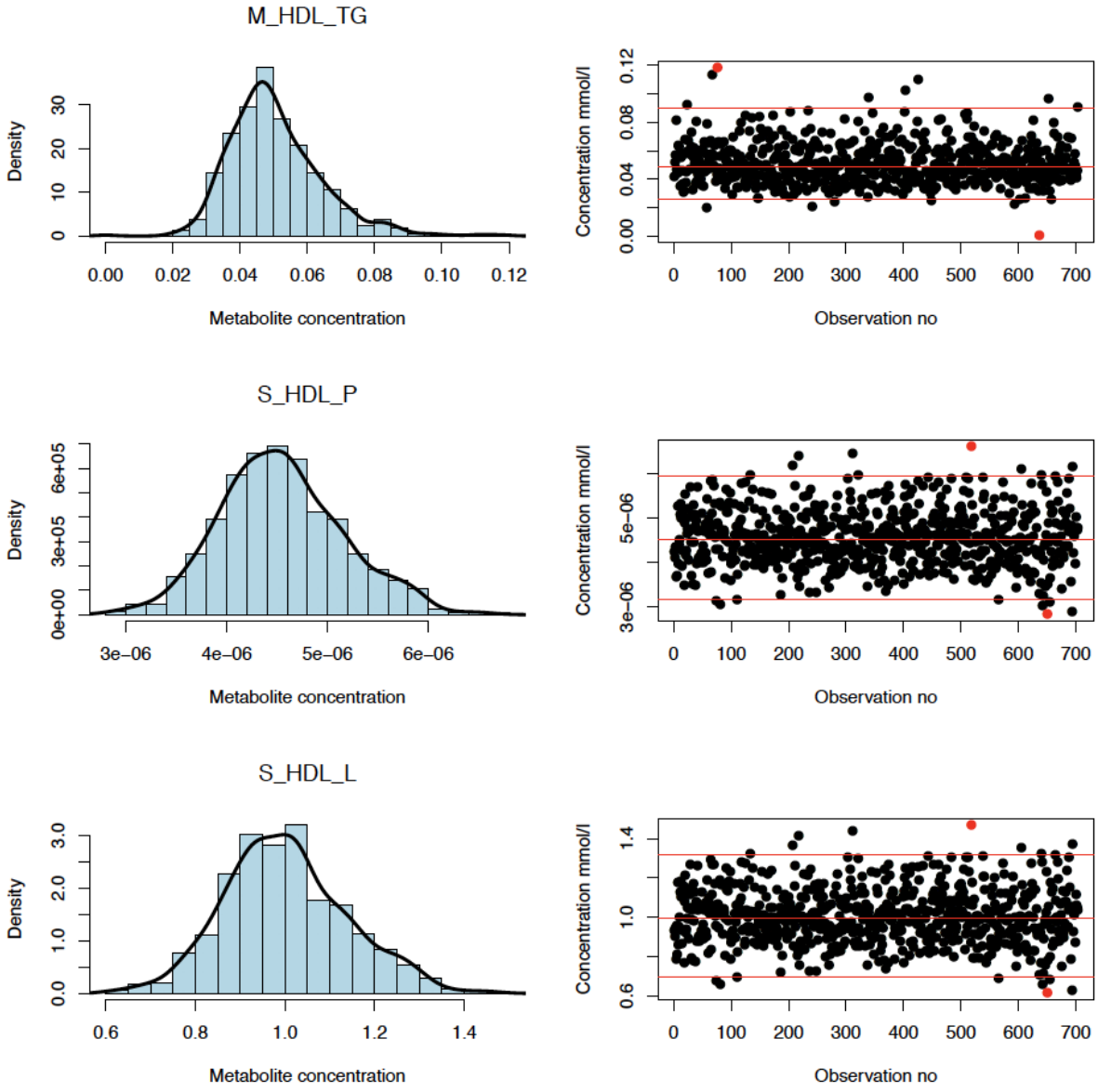
D4 continued.



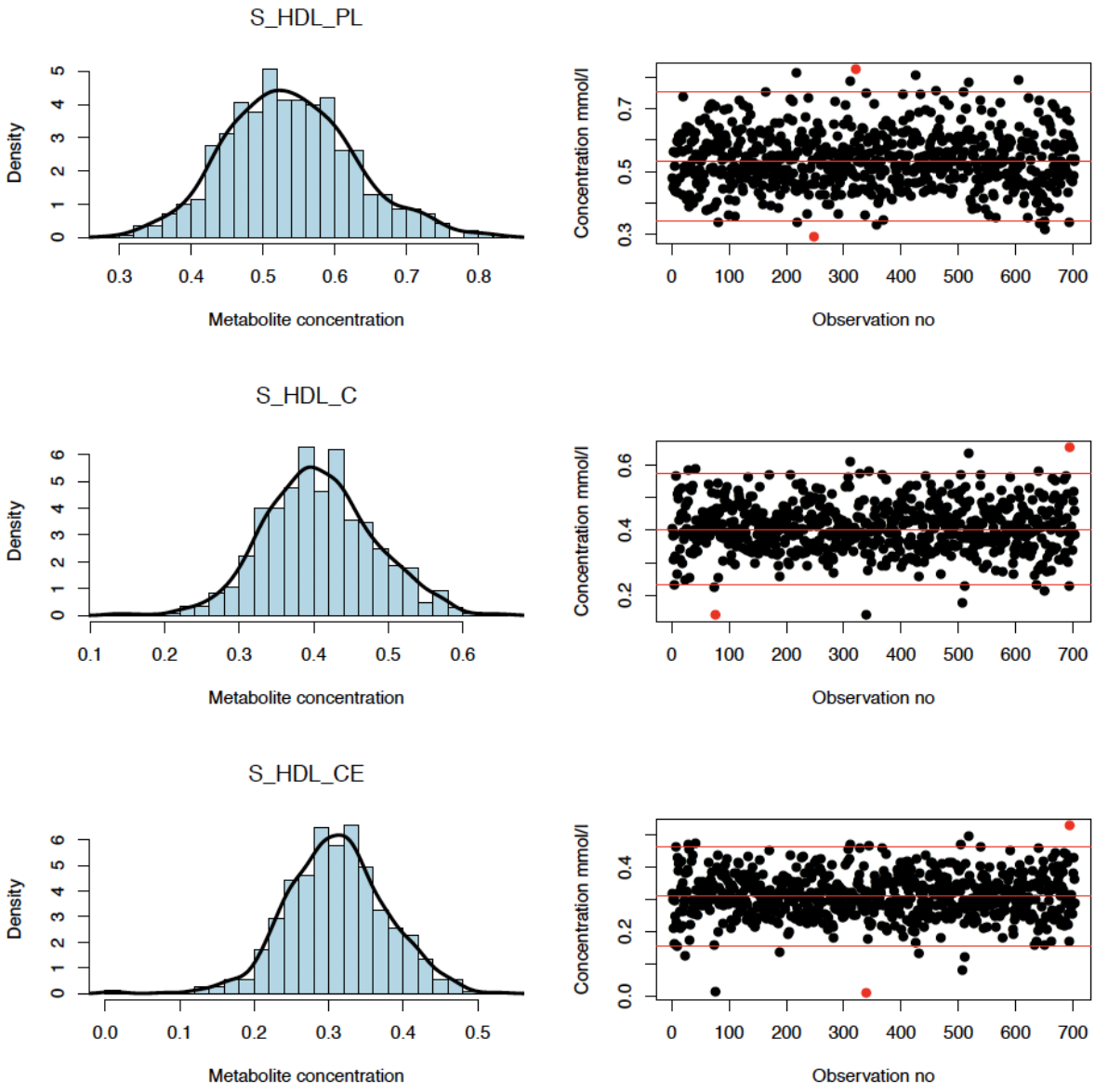
D4 continued.



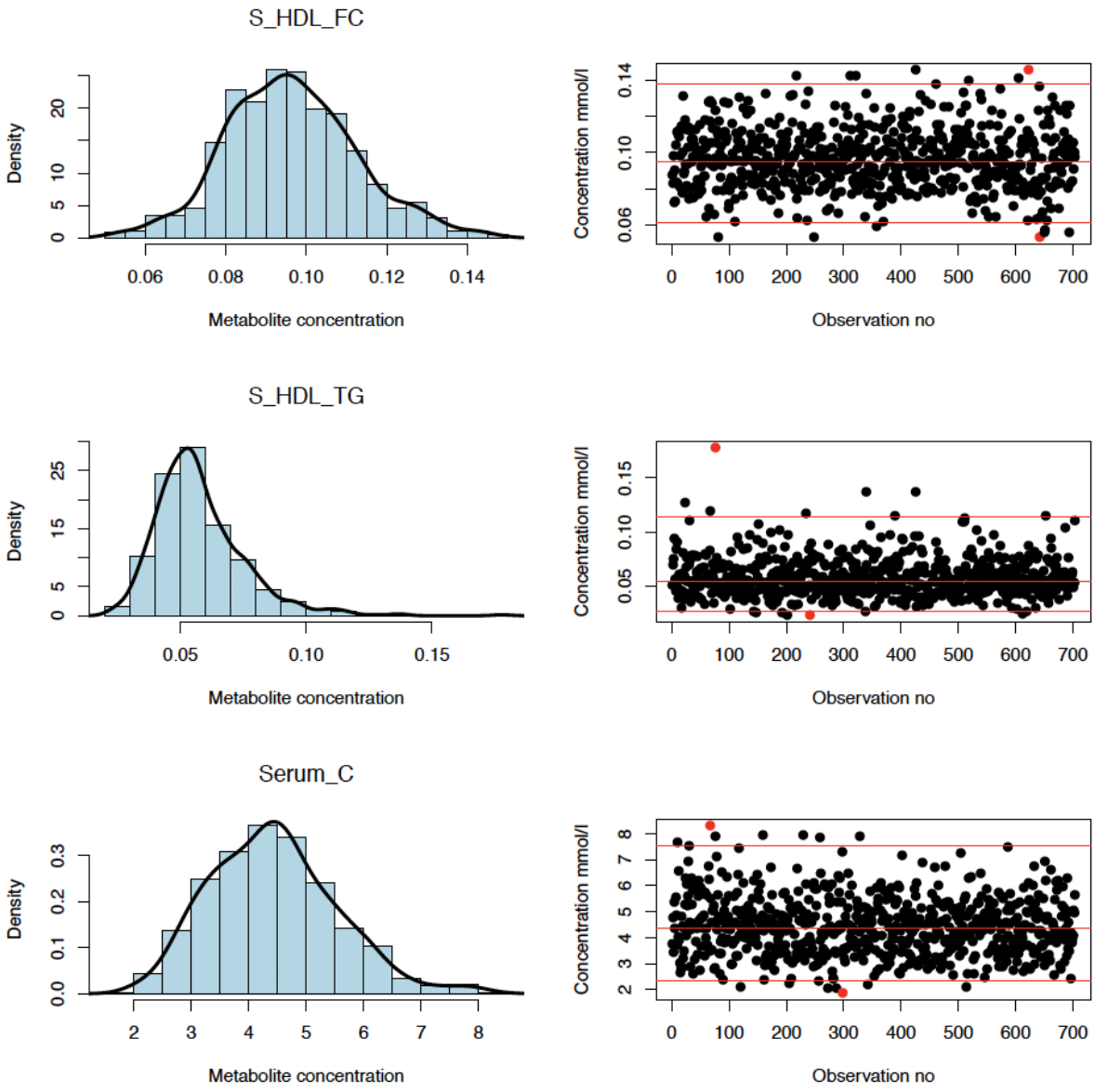
D4 continued.



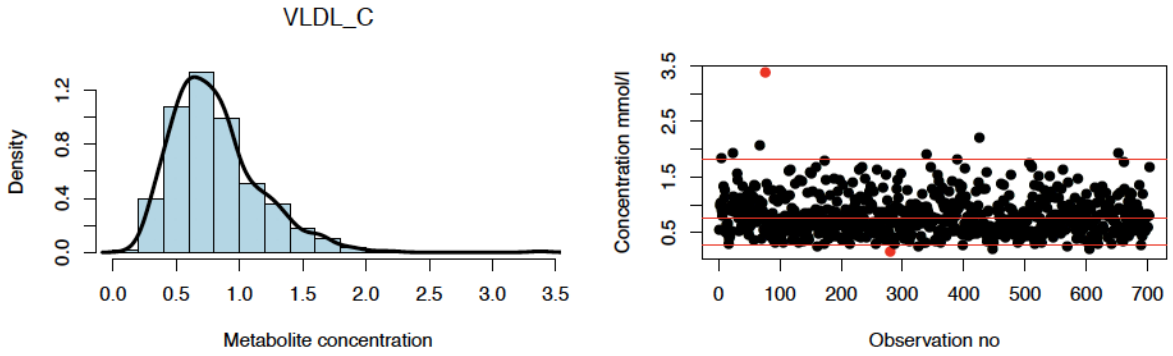
D4 continued.

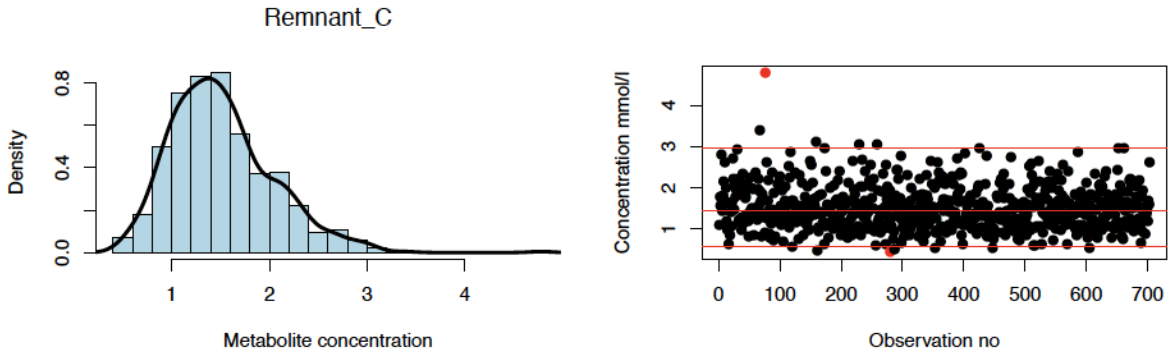


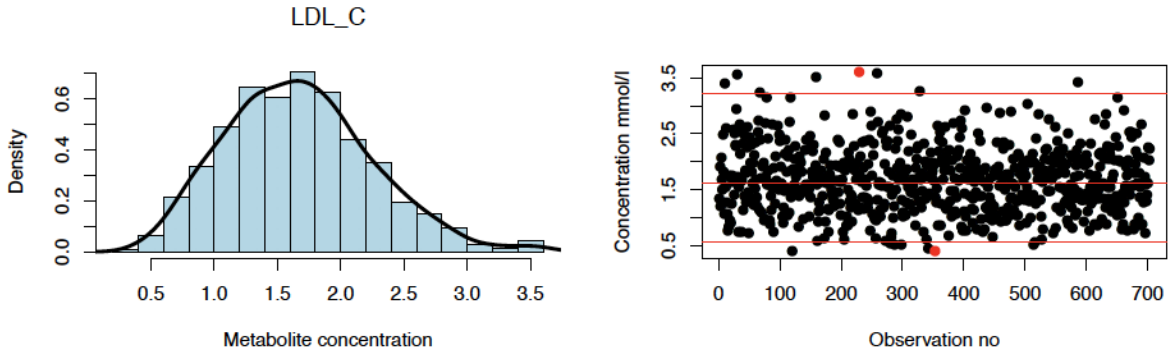
D4 continued.



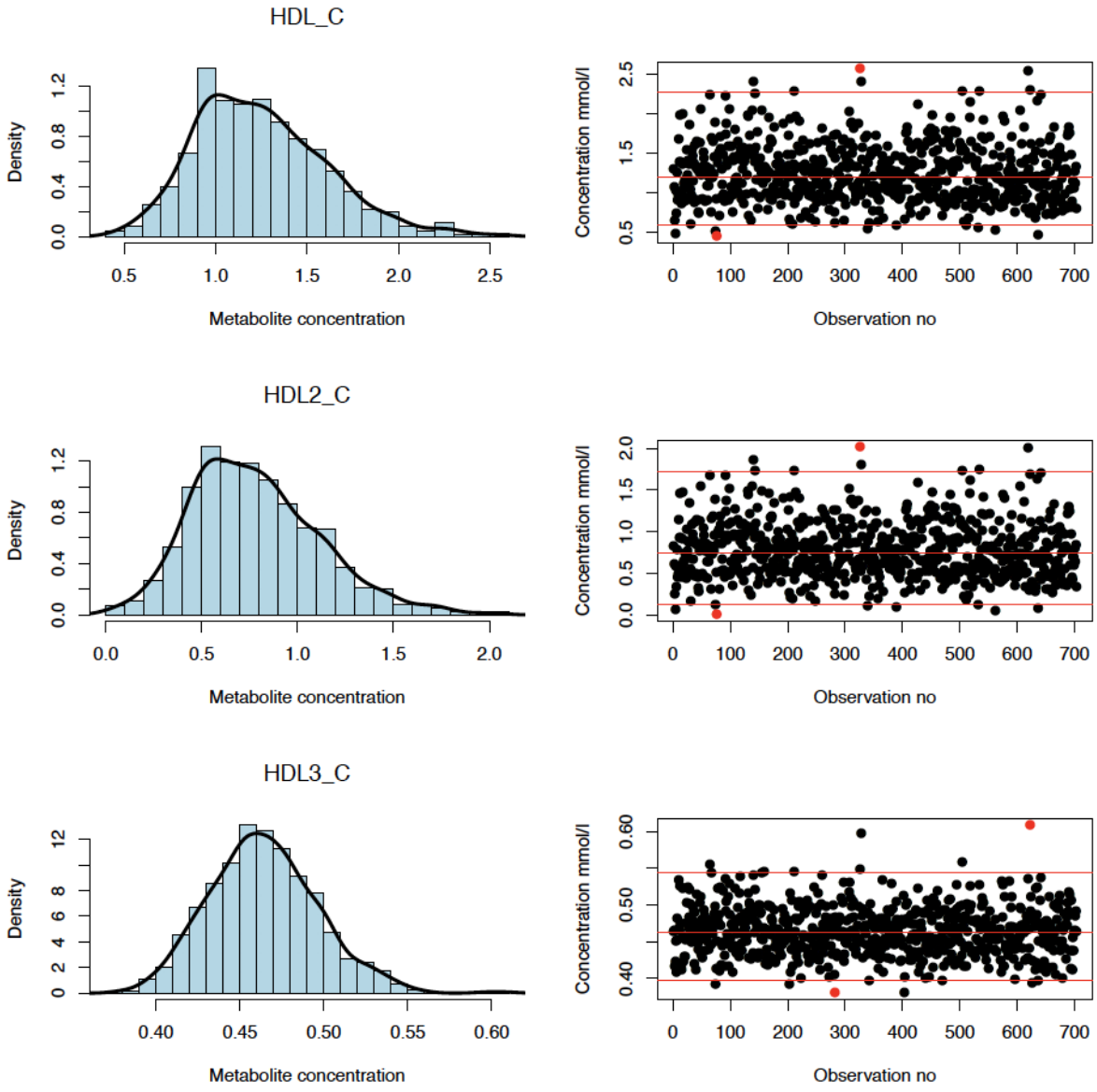
D4 continued.



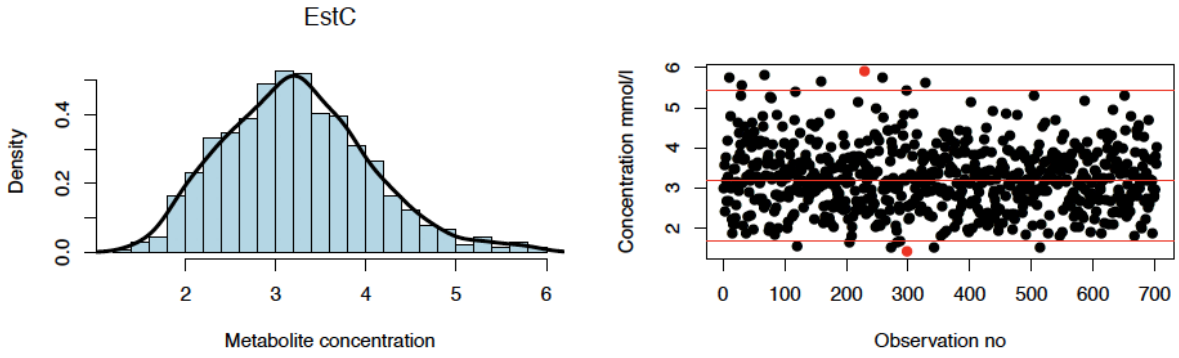


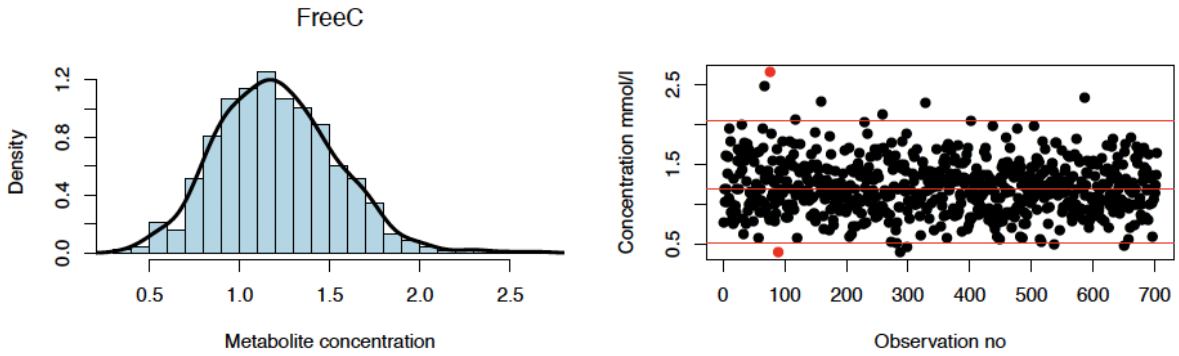


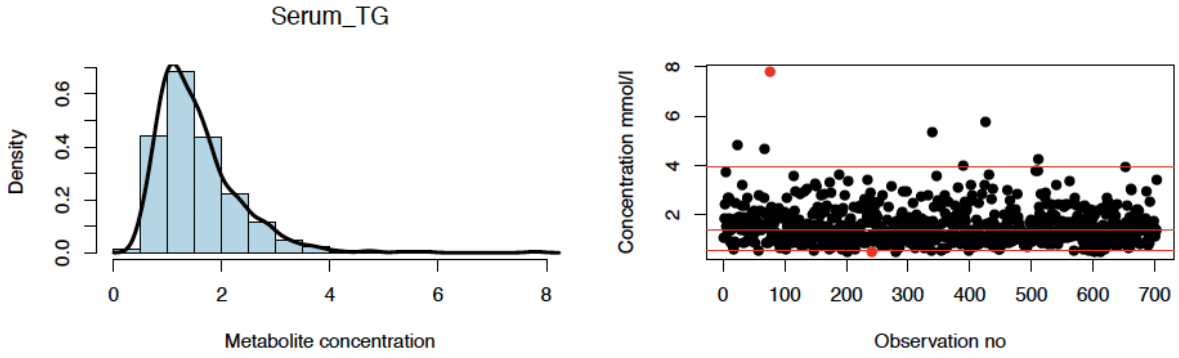
D4 continued.



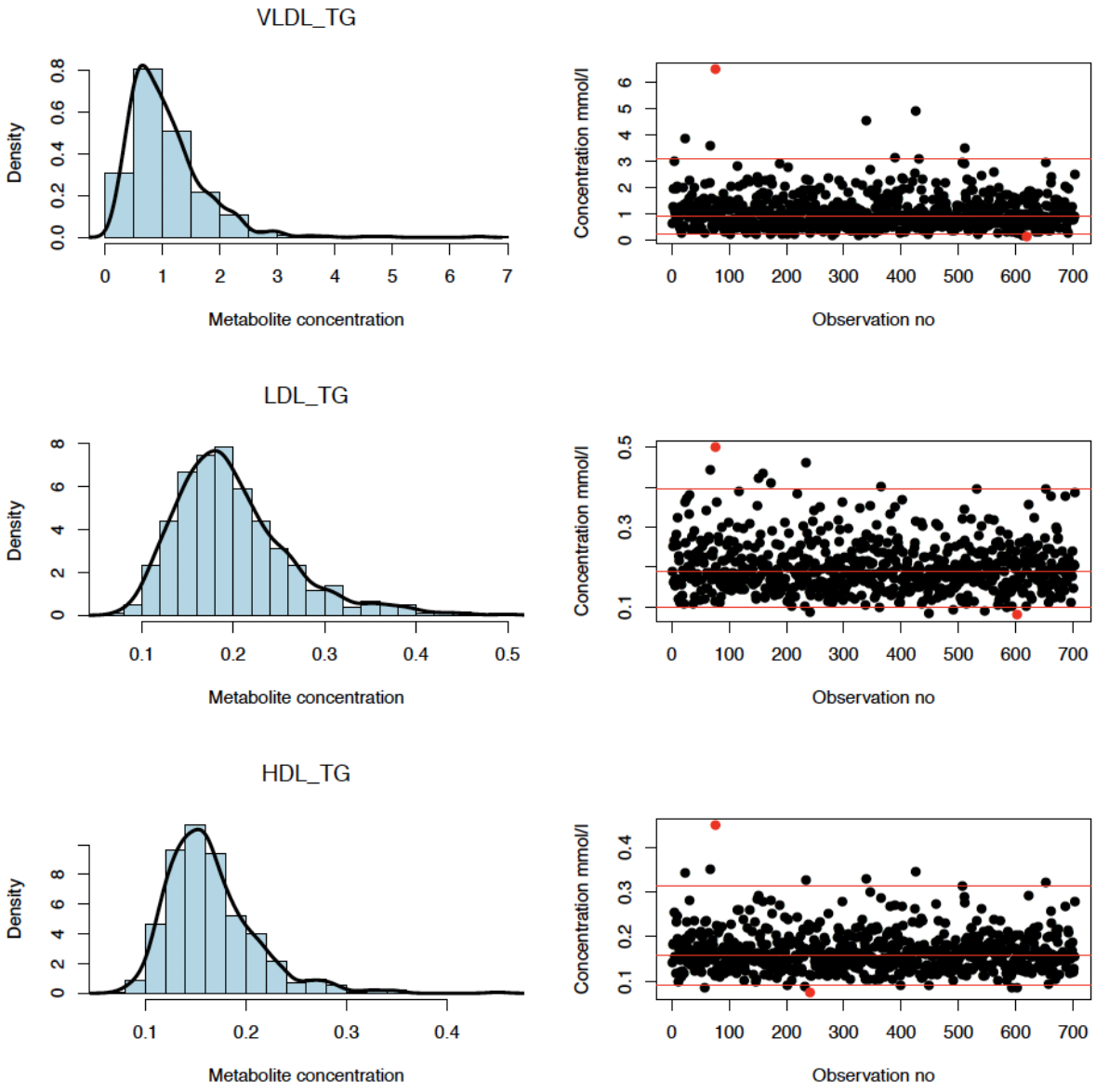
D4 continued.



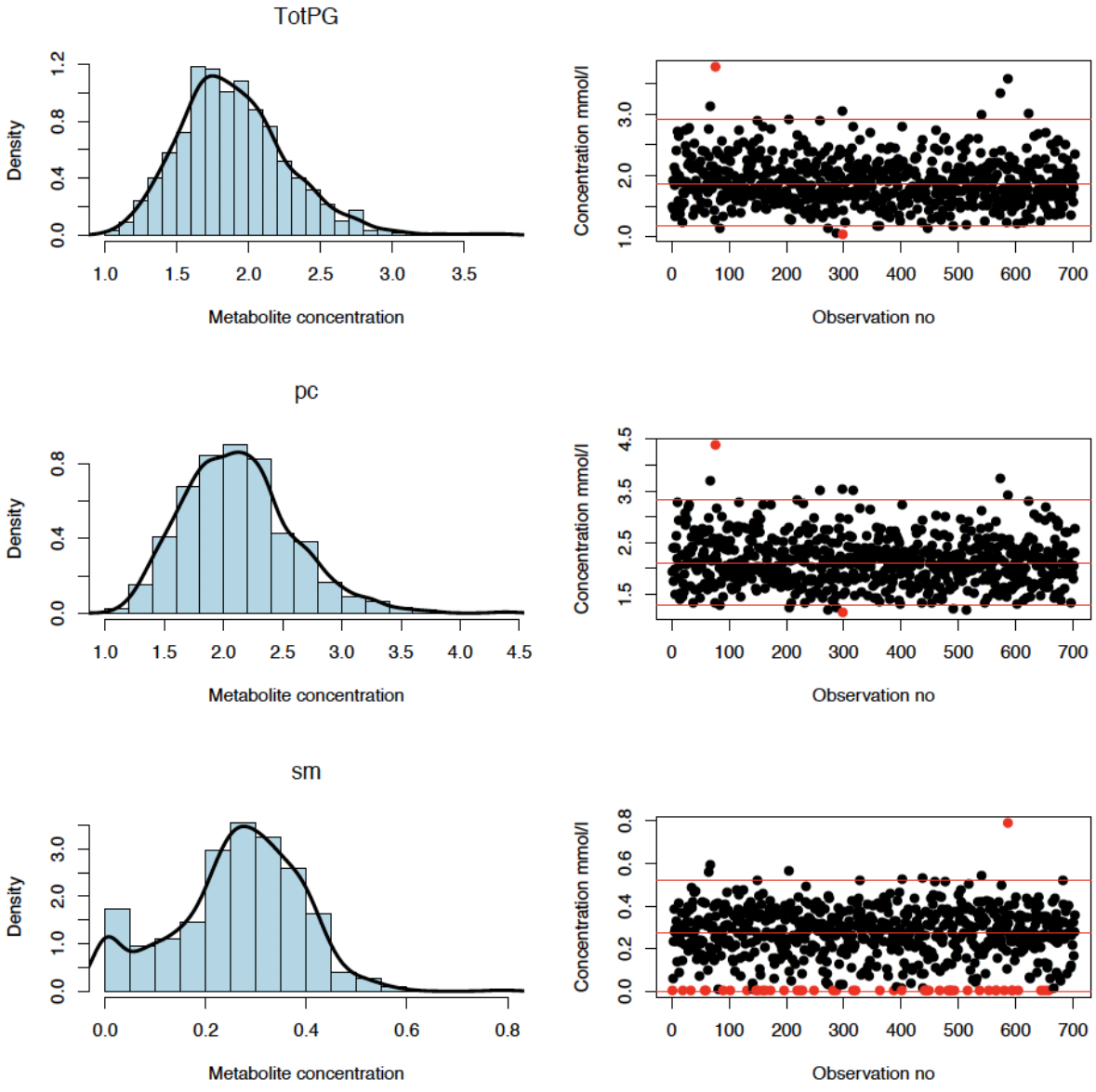




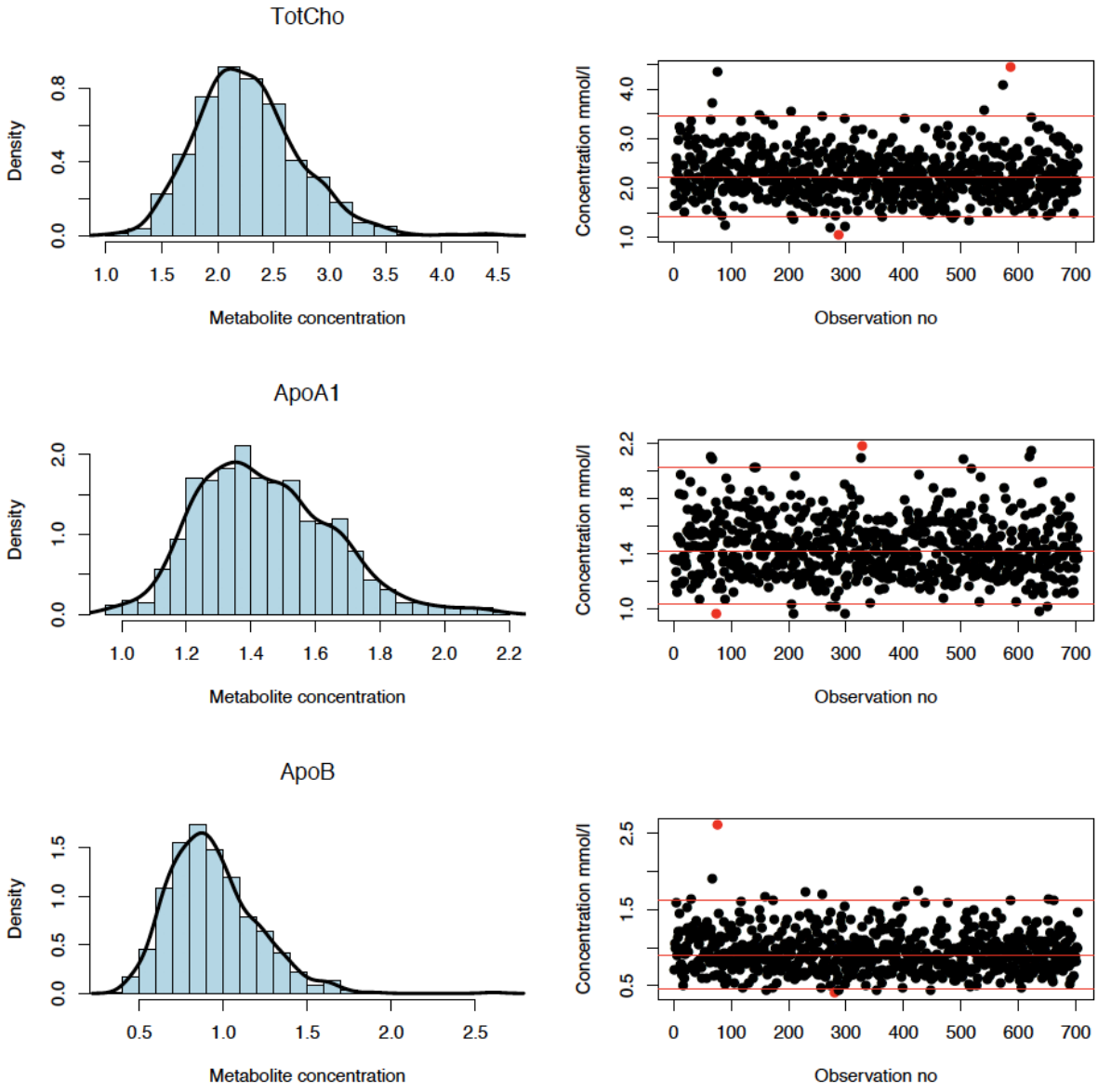
D4 continued.



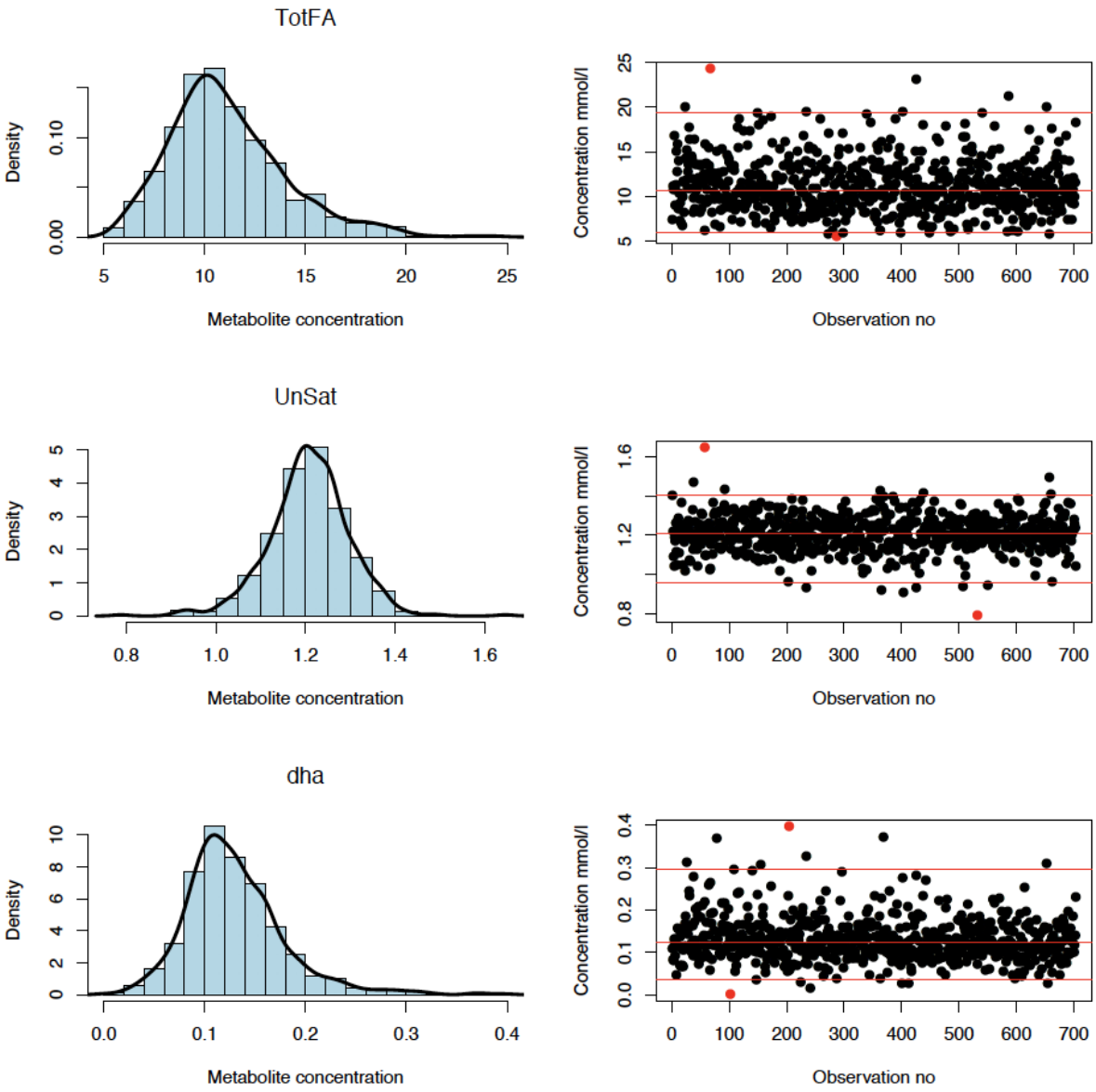
D4 continued.



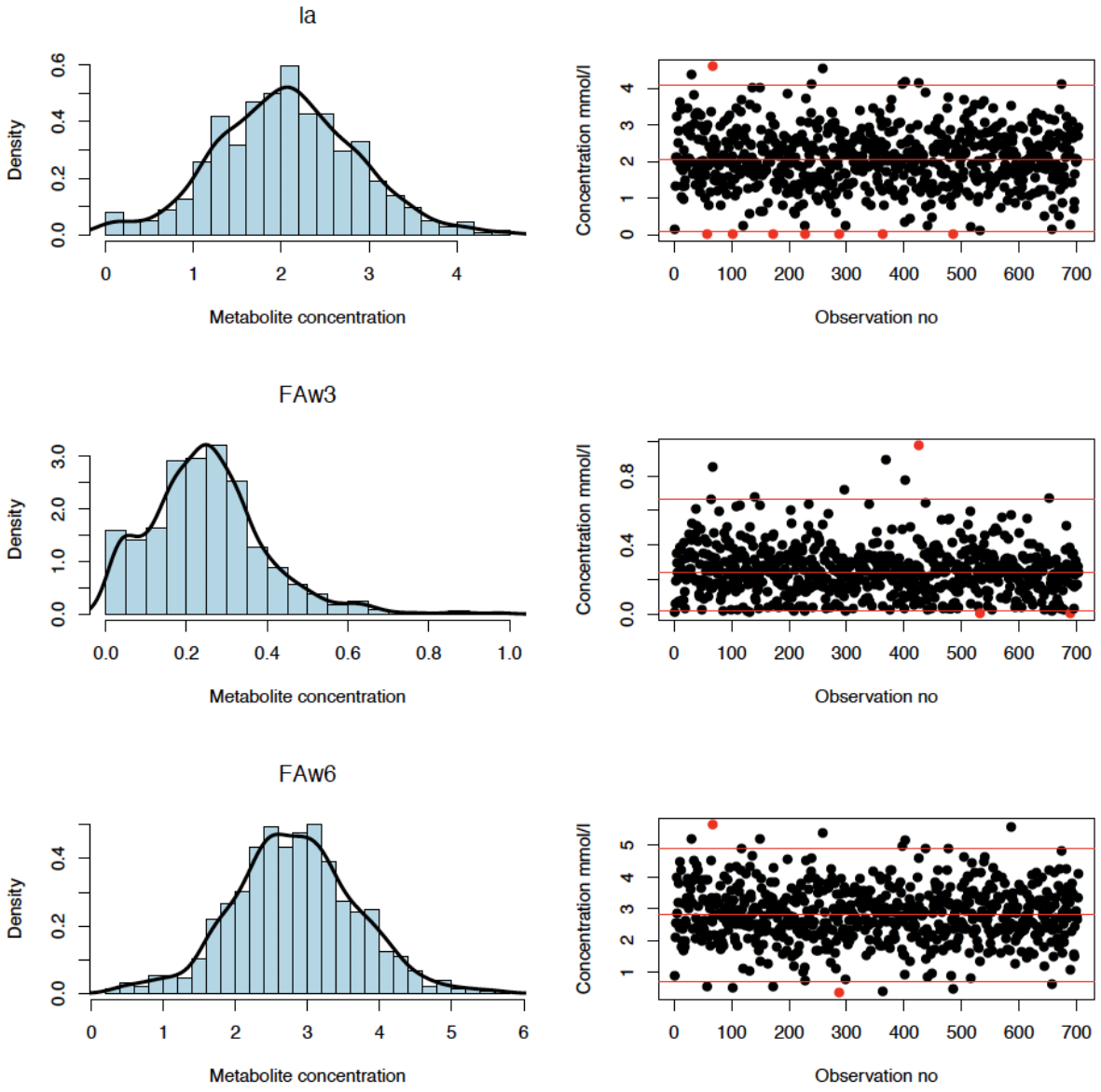
D4 continued.



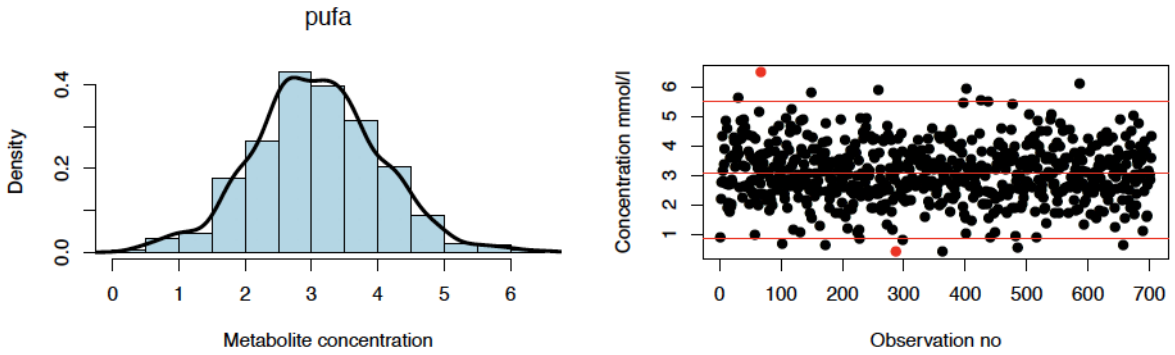
D4 continued.

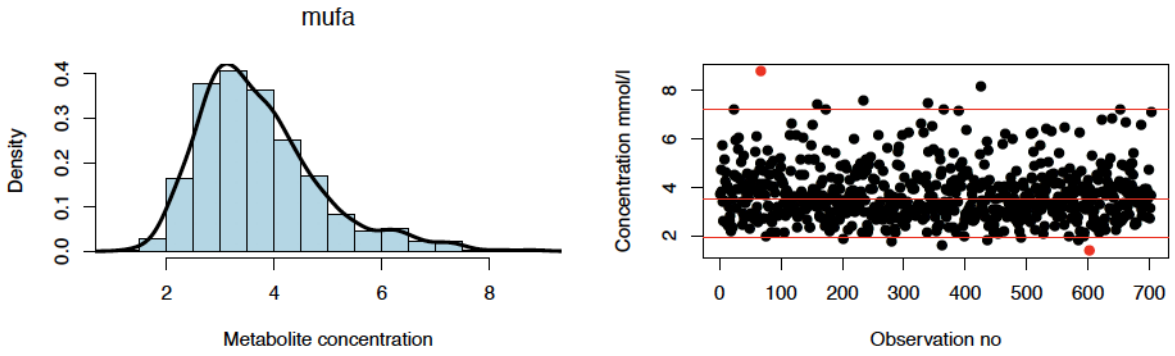


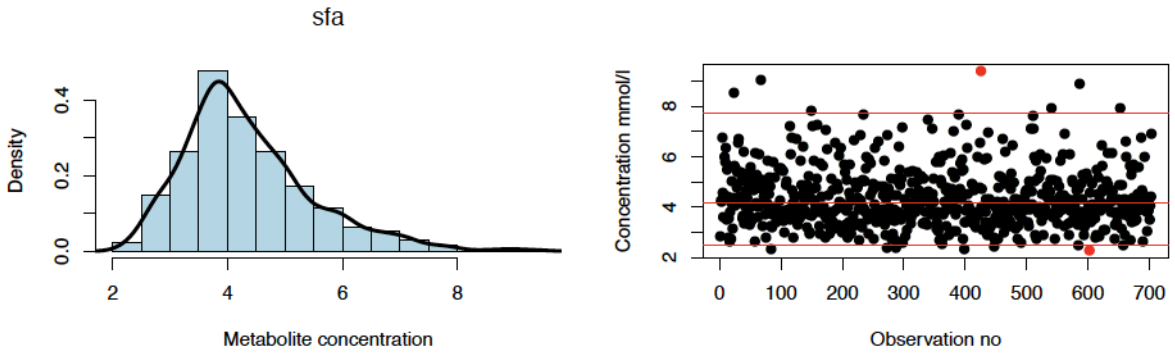
D4 continued.



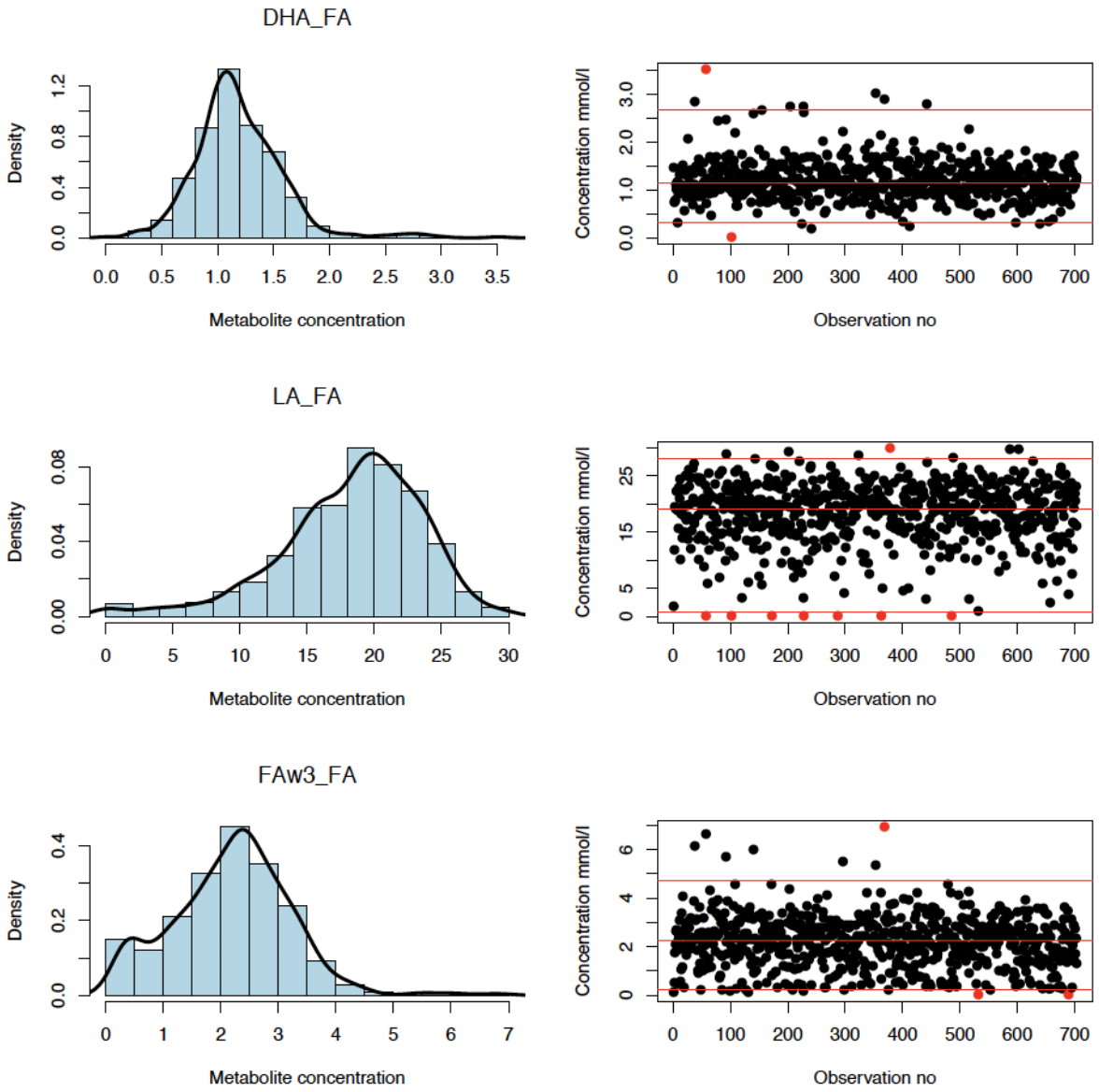
D4 continued.



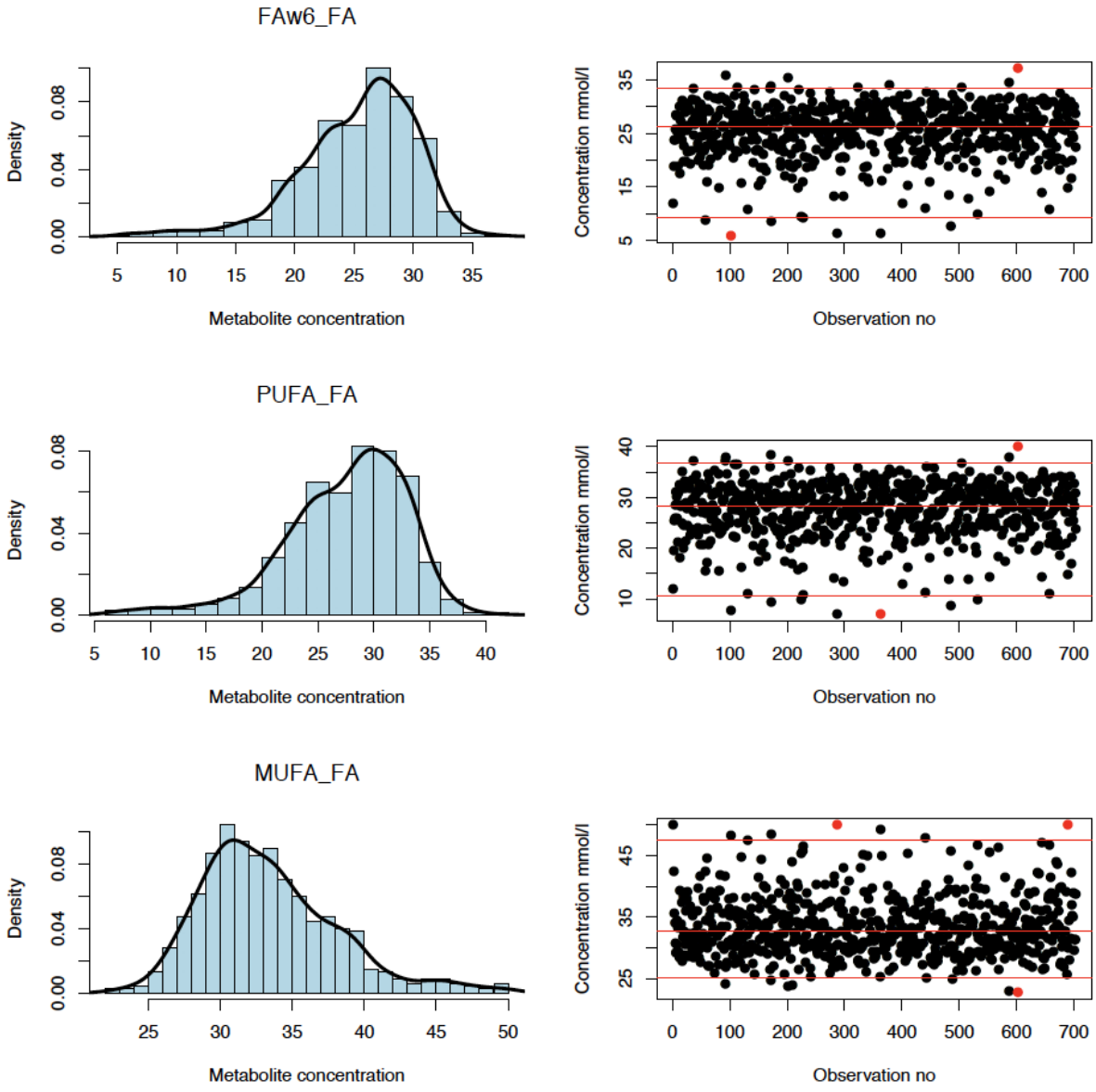




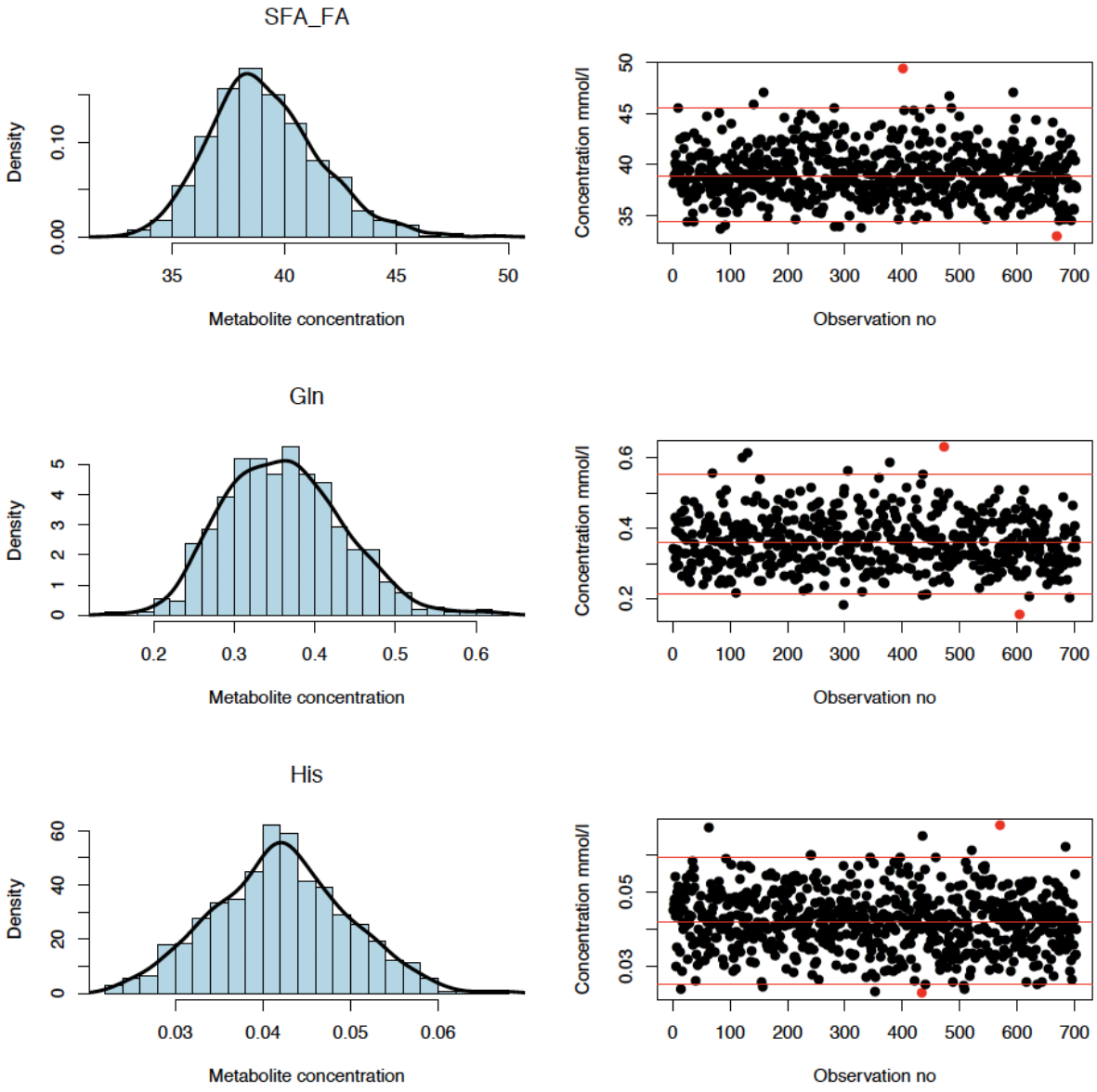
D4 continued.



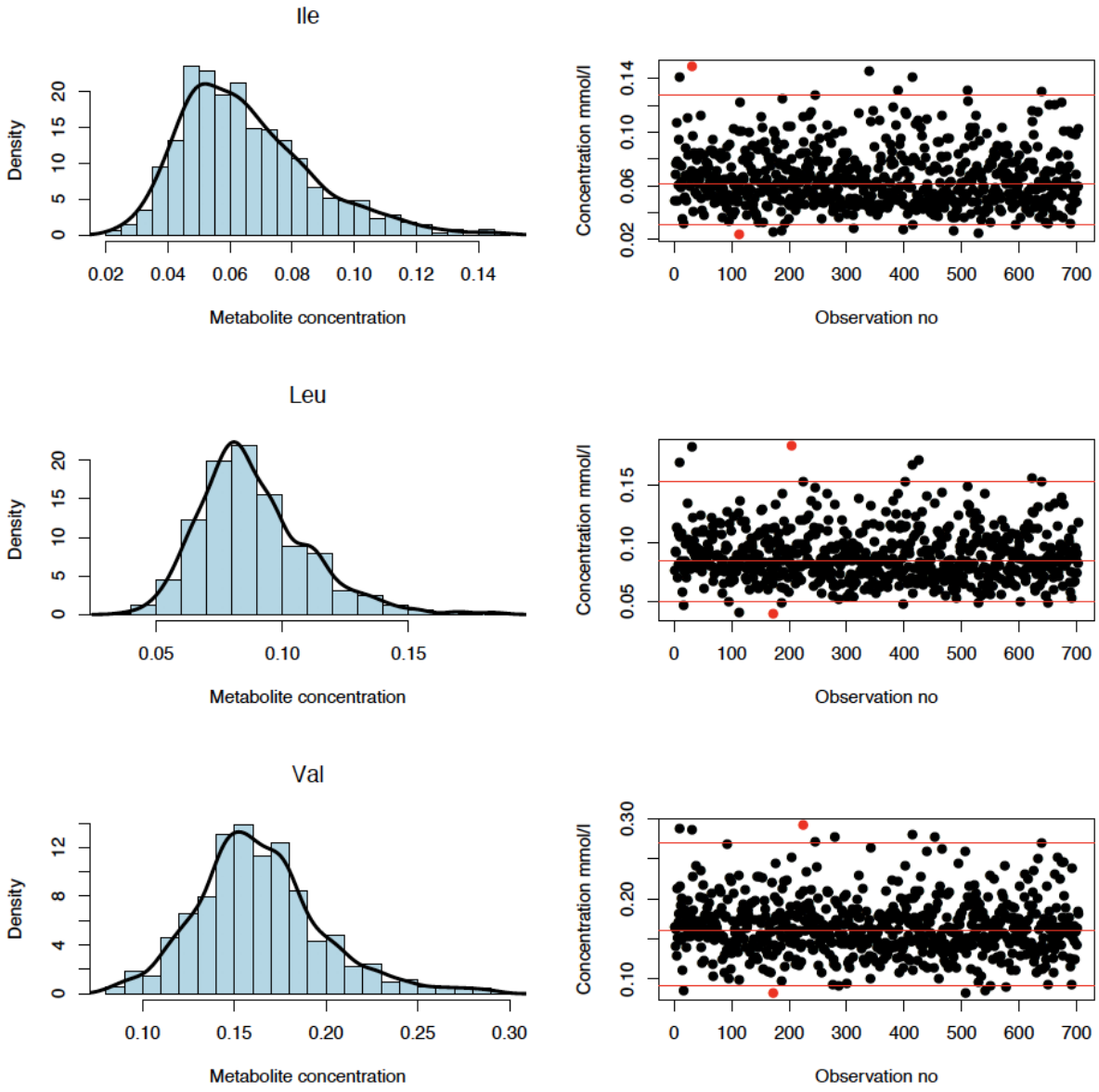
D4 continued.



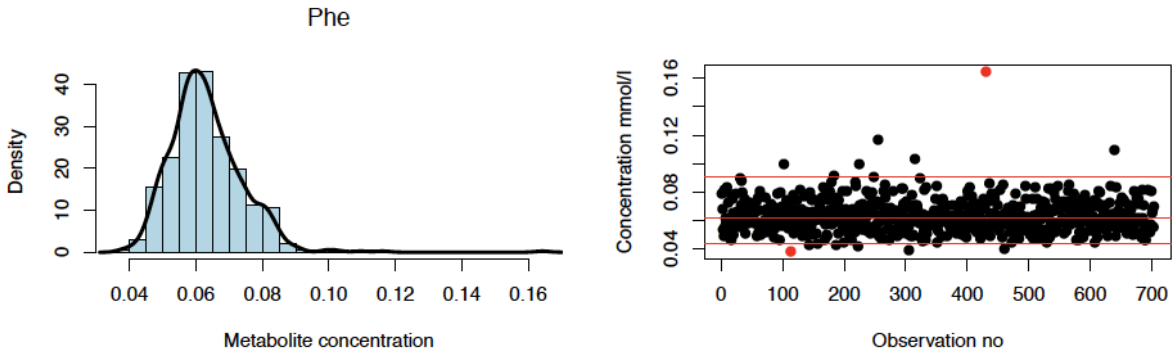
D4 continued.

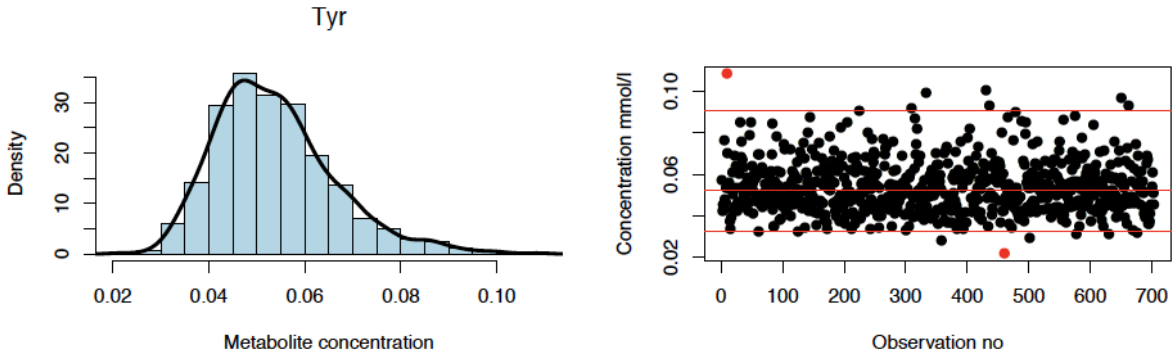


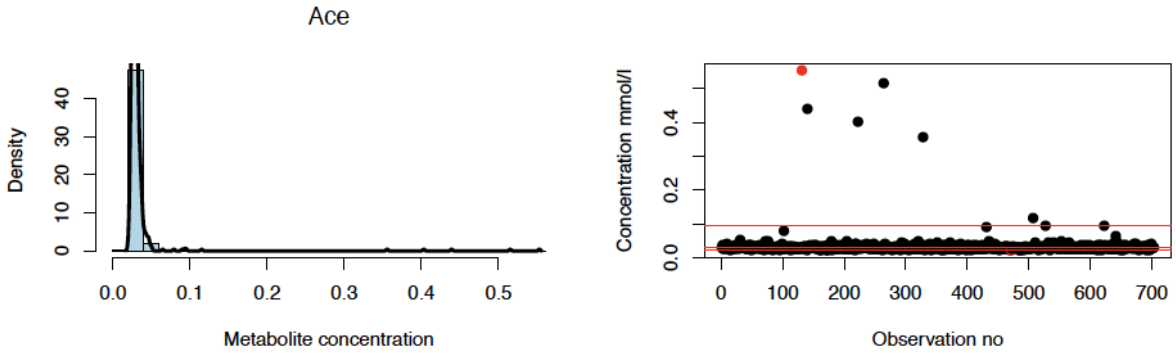
D4 continued.



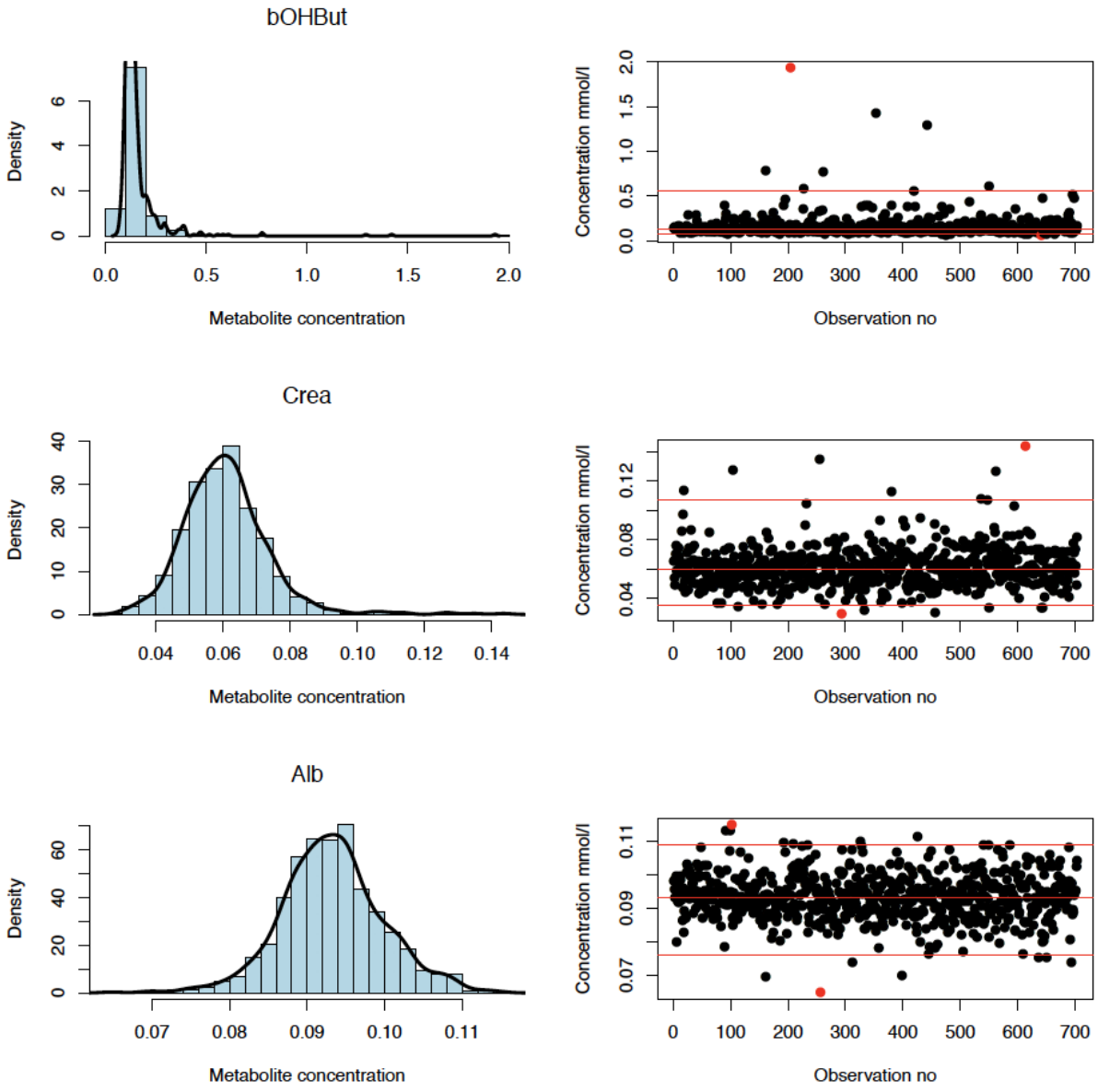
D4 continued.



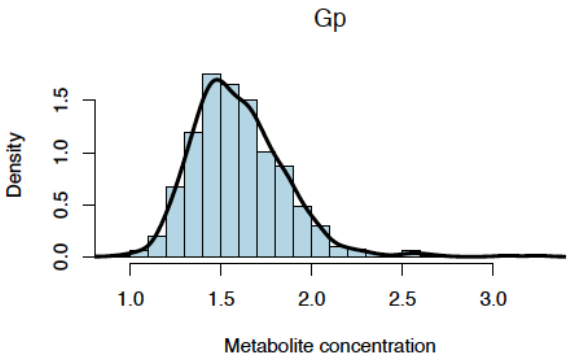


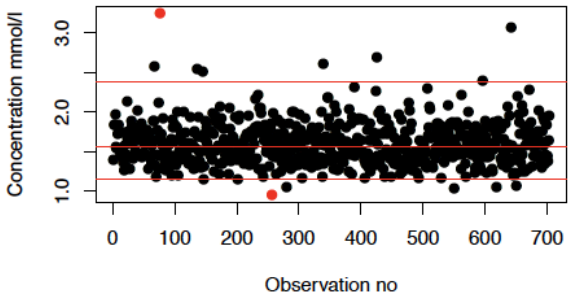


D4 continued.



D4 continued.





D5: Standardised mean difference in circulating metabolic trait concentrations for HPV (+) vs. HPV (-) OPCs.

Metabolite	N	b	SE	Lower CI	Upper CI	*p*-value*
L_HDL_CE	703	-0.503	0.083	-0.666	-0.340	2.14E-09
L_HDL_C	703	-0.501	0.083	-0.664	-0.338	2.55E-09
L_HDL_FC	703	-0.493	0.083	-0.656	-0.329	4.99E-09
L_HDL_L	703	-0.471	0.084	-0.635	-0.307	2.52E-08
L_HDL_P	703	-0.468	0.084	-0.632	-0.304	3.12E-08
XL_HDL_PL	703	-0.464	0.084	-0.628	-0.300	3.88E-08
S_VLDL_CE	703	0.474	0.086	0.305	0.642	4.86E-08
S_VLDL_C	703	0.458	0.085	0.291	0.625	1.03E-07
Crea	685	0.406	0.077	0.255	0.556	1.77E-07
L_HDL_PL	703	-0.427	0.084	-0.593	-0.262	5.06E-07
ApoB	703	0.435	0.086	0.266	0.604	5.78E-07
Val	702	0.413	0.082	0.252	0.573	6.14E-07
His	703	0.423	0.085	0.256	0.591	8.86E-07
XL_HDL_FC	703	-0.418	0.084	-0.584	-0.253	8.99E-07
M_LDL_CE	703	0.424	0.086	0.255	0.592	1.02E-06
S_VLDL_L	703	0.413	0.084	0.248	0.578	1.12E-06
M_LDL_C	703	0.419	0.086	0.250	0.588	1.34E-06
S_LDL_C	703	0.417	0.086	0.249	0.586	1.37E-06
S_LDL_CE	703	0.417	0.086	0.249	0.586	1.41E-06
XL_HDL_P	703	-0.410	0.085	-0.576	-0.244	1.52E-06
XL_HDL_L	703	-0.409	0.085	-0.575	-0.243	1.59E-06
S_VLDL_P	703	0.406	0.084	0.241	0.571	1.66E-06
S_VLDL_PL	703	0.403	0.083	0.239	0.567	1.69E-06
S_LDL_L	703	0.410	0.086	0.242	0.579	2.11E-06
S_LDL_P	703	0.407	0.086	0.239	0.575	2.54E-06
M_LDL_PL	703	0.408	0.086	0.239	0.577	2.57E-06
S_LDL_FC	703	0.404	0.085	0.236	0.571	2.71E-06
bOHBut	693	-0.299	0.063	-0.423	-0.175	2.71E-06
S_VLDL_FC	703	0.392	0.084	0.227	0.557	3.61E-06
Remnant_C	703	0.404	0.087	0.234	0.574	3.67E-06
M_LDL_L	703	0.402	0.086	0.232	0.571	3.78E-06
S_LDL_PL	703	0.396	0.085	0.228	0.564	4.35E-06
LDL_C	703	0.398	0.086	0.228	0.567	4.86E-06
Ile	700	0.382	0.083	0.218	0.545	5.26E-06
VLDL_C	703	0.392	0.086	0.224	0.560	5.33E-06
M_LDL_P	703	0.395	0.086	0.226	0.564	5.46E-06
M_LDL_FC	703	0.393	0.086	0.225	0.562	5.50E-06
L_LDL_CE	703	0.395	0.087	0.224	0.565	6.18E-06
la	680	0.400	0.089	0.224	0.575	8.87E-06
XS_VLDL_FC	703	0.383	0.086	0.214	0.553	1.01E-05
M_VLDL_CE	703	0.381	0.086	0.212	0.549	1.05E-05
L_LDL_C	703	0.375	0.087	0.205	0.545	1.75E-05

D5 continued.

Metabolite	N	b	SE	Lower CI	Upper CI	*p*-value*
L_HDL_TG	703	-0.365	0.085	-0.533	-0.198	2.10E-05
D4 continued.						
XS_VLDL_PL	703	0.369	0.087	0.199	0.540	2.39E-05
S_VLDL_TG	703	0.355	0.084	0.190	0.520	2.61E-05
L_LDL_PL	703	0.368	0.087	0.197	0.538	2.66E-05
pufa	680	0.377	0.090	0.201	0.553	2.89E-05
M_VLDL_C	703	0.359	0.085	0.191	0.526	2.95E-05
M_VLDL_PL	703	0.354	0.085	0.188	0.521	3.34E-05
M_VLDL_P	703	0.354	0.085	0.187	0.521	3.41E-05
M_VLDL_L	703	0.354	0.085	0.187	0.520	3.50E-05
M_VLDL_TG	703	0.348	0.085	0.182	0.515	4.54E-05
FAw6	680	0.367	0.090	0.191	0.543	4.61E-05
L_LDL_L	703	0.355	0.087	0.184	0.526	4.91E-05
XS_VLDL_L	703	0.353	0.086	0.183	0.522	4.96E-05
XS_VLDL_P	703	0.349	0.086	0.180	0.518	5.78E-05
L_LDL_P	703	0.351	0.087	0.181	0.522	5.93E-05
IDL_CE	703	0.351	0.087	0.180	0.523	6.22E-05
HDL2_C	703	-0.336	0.084	-0.502	-0.170	7.58E-05
XL_HDL_C	703	-0.334	0.085	-0.501	-0.166	1.01E-04
Leu	700	0.330	0.085	0.164	0.496	1.07E-04
VLDL_TG	703	0.329	0.085	0.163	0.496	1.11E-04
XS_VLDL_C	703	0.334	0.087	0.164	0.505	1.27E-04
M_VLDL_FC	703	0.326	0.085	0.160	0.493	1.33E-04
IDL_C	703	0.332	0.087	0.161	0.504	1.52E-04
sfa	680	0.331	0.088	0.158	0.504	1.86E-04
L_VLDL_CE	703	0.318	0.085	0.152	0.485	1.87E-04
HDL_C	703	-0.316	0.084	-0.482	-0.151	1.90E-04
MUFA_FA	680	-0.316	0.086	-0.484	-0.148	2.38E-04
FAw3	680	0.329	0.090	0.152	0.506	2.73E-04
L_VLDL_TG	703	0.309	0.085	0.142	0.475	3.03E-04
L_VLDL_P	703	0.308	0.085	0.142	0.475	3.04E-04
EstC	695	0.320	0.088	0.147	0.493	3.05E-04
L_VLDL_L	703	0.307	0.085	0.141	0.474	3.18E-04
L_VLDL_PL	703	0.307	0.085	0.140	0.473	3.30E-04
IDL_L	703	0.313	0.087	0.142	0.485	3.61E-04
L_VLDL_C	703	0.304	0.085	0.137	0.471	3.72E-04
IDL_PL	703	0.311	0.087	0.139	0.482	4.02E-04
IDL_P	703	0.310	0.087	0.138	0.482	4.14E-04
XS_VLDL_CE	703	0.308	0.087	0.138	0.479	4.16E-04
L_LDL_FC	703	0.306	0.087	0.136	0.477	4.55E-04
Serum_TG	703	0.295	0.085	0.128	0.461	5.30E-04
XL_HDL_CE	703	-0.295	0.086	-0.463	-0.127	6.01E-04
XXL_VLDL_CE	703	0.294	0.086	0.125	0.463	6.74E-04

D5 continued.

Metabolite	N	b	SE	Lower CI	Upper CI	*p*-value*
Serum_C	703	0.297	0.088	0.125	0.469	7.50E-04
XL_VLDL_CE	703	0.288	0.085	0.121	0.456	7.71E-04
L_VLDL_FC	703	0.288	0.085	0.120	0.455	7.78E-04
XL_VLDL_PL	703	0.289	0.086	0.121	0.456	7.84E-04
XL_VLDL_TG	703	0.288	0.085	0.120	0.455	7.93E-04
XL_VLDL_P	703	0.288	0.085	0.120	0.455	7.94E-04
XL_VLDL_L	703	0.287	0.085	0.120	0.455	8.13E-04
TotFA	680	0.297	0.088	0.123	0.471	8.35E-04
XL_VLDL_C	703	0.283	0.086	0.115	0.451	9.88E-04
XXL_VLDL_P	703	0.282	0.086	0.113	0.451	0.001
XXL_VLDL_TG	703	0.282	0.086	0.113	0.451	0.001
XXL_VLDL_L	703	0.281	0.086	0.112	0.450	0.001
XXL_VLDL_C	703	0.280	0.086	0.111	0.448	0.001
S_HDL_CE	703	0.269	0.083	0.105	0.432	0.001
XL_VLDL_FC	703	0.276	0.086	0.108	0.444	0.001
XXL_VLDL_PL	703	0.275	0.086	0.106	0.444	0.001
XS_VLDL_TG	703	0.266	0.084	0.101	0.430	0.002
IDL_FC	703	0.270	0.087	0.099	0.442	0.002
LA_FA	680	0.265	0.086	0.097	0.434	0.002
XXL_VLDL_FC	703	0.255	0.086	0.086	0.423	0.003
Ace	696	-0.150	0.051	-0.251	-0.050	0.003
M_HDL_TG	703	0.234	0.083	0.072	0.396	0.005
S_HDL_C	703	0.232	0.083	0.070	0.394	0.005
FreeC	695	0.250	0.089	0.074	0.425	0.005
FAw3_FA	680	0.239	0.087	0.069	0.410	0.006
PUFA_FA	680	0.210	0.087	0.040	0.381	0.016
S_HDL_TG	703	0.192	0.084	0.027	0.356	0.022
S_LDL_TG	703	0.182	0.085	0.016	0.348	0.032
FAw6_FA	680	0.182	0.087	0.011	0.353	0.037
DHA_FA	680	-0.155	0.081	-0.315	0.004	0.056
SFA_FA	680	0.151	0.082	-0.011	0.312	0.067
UnSat	680	0.158	0.087	-0.013	0.328	0.070
M_HDL_PL	703	-0.142	0.084	-0.307	0.023	0.091
M_HDL_FC	703	-0.142	0.086	-0.311	0.026	0.098
Alb	703	0.134	0.083	-0.029	0.297	0.107
M_HDL_C	703	-0.128	0.086	-0.297	0.041	0.139
M_HDL_CE	703	-0.123	0.086	-0.293	0.046	0.153
pc	695	0.124	0.088	-0.049	0.298	0.160
M_HDL_L	703	-0.119	0.085	-0.286	0.048	0.162
ApoA1	703	-0.119	0.086	-0.288	0.051	0.170
Phe	700	0.115	0.084	-0.050	0.280	0.173
S_HDL_P	703	0.111	0.082	-0.050	0.271	0.176
M_HDL_P	703	-0.112	0.085	-0.279	0.054	0.185

D5 continued.

Metabolite	N	b	SE	Lower CI	Upper CI	*p*-value*
S_HDL_L	703	0.108	0.082	-0.052	0.269	0.187
IDL_TG	703	0.110	0.085	-0.056	0.277	0.193
mufa	680	0.108	0.085	-0.060	0.275	0.207
LDL_TG	703	0.074	0.085	-0.093	0.241	0.384
sm	695	0.072	0.087	-0.098	0.243	0.405
TotCho	695	0.069	0.090	-0.107	0.245	0.442
S_HDL_PL	703	-0.063	0.084	-0.228	0.102	0.454
dha	680	0.059	0.086	-0.109	0.227	0.489
L_LDL_TG	703	0.054	0.085	-0.113	0.222	0.525
TotPG	695	0.055	0.088	-0.119	0.228	0.537
Gp	703	-0.049	0.086	-0.217	0.120	0.572
S_HDL_FC	703	-0.042	0.084	-0.206	0.122	0.617
HDL3_C	703	-0.040	0.082	-0.201	0.121	0.624
Gln	548	-0.045	0.095	-0.231	0.142	0.639
Tyr	702	0.037	0.080	-0.121	0.195	0.644
HDL_TG	703	0.034	0.085	-0.133	0.200	0.692
M_LDL_TG	703	0.032	0.085	-0.135	0.199	0.710
XL_HDL_TG	703	-0.004	0.086	-0.172	0.164	0.961

Abbreviations: ***b****, the regression coefficient from the linear regression model;* ***CI****, 95% confidence interval;* ***N****, sample number.*

**P-value for difference.*

D6: Summary of Cox PH regression results, showing only those metabolites that reached the threshold for multiple testing (n=703).

Metabolite	N	b	Lower CI	Upper CI	*p*-value
Model 1					
Crea	685	0.54	0.41	0.695	2.69E-06
Ace	696	1.38	1.20	1.592	7.56E-06
His	703	0.68	0.57	0.821	5.34E-05
L_HDL_CE	703	1.39	1.18	1.631	6.51E-05
L_HDL_C	703	1.38	1.17	1.621	8.96E-05
FAw3_FA	680	0.70	0.58	0.842	1.65E-04
FAw3	680	0.68	0.56	0.831	1.66E-04
L_HDL_TG	703	1.39	1.17	1.642	1.66E-04
L_HDL_P	703	1.37	1.16	1.617	1.78E-04
L_HDL_L	703	1.37	1.16	1.615	1.81E-04
L_HDL_FC	703	1.35	1.15	1.589	2.49E-04
S_LDL_CE	703	0.72	0.60	0.870	0.001
S_LDL_C	703	0.72	0.60	0.871	0.001
M_LDL_CE	703	0.73	0.61	0.874	0.001
L_HDL_PL	703	1.34	1.13	1.585	0.001
XL_HDL_PL	703	1.32	1.13	1.559	0.001
M_LDL_C	703	0.73	0.61	0.877	0.001
S_LDL_FC	703	0.73	0.61	0.881	0.001
S_LDL_L	703	0.73	0.61	0.882	0.001
S_LDL_P	703	0.73	0.61	0.886	0.001
M_LDL_FC	703	0.74	0.62	0.893	0.002
M_LDL_L	703	0.74	0.62	0.894	0.002
LDL_C	703	0.75	0.62	0.898	0.002
S_HDL_CE	703	0.75	0.63	0.901	0.002
M_LDL_P	703	0.75	0.62	0.899	0.002
S_LDL_PL	703	0.75	0.62	0.901	0.002
M_LDL_PL	703	0.75	0.62	0.904	0.003
S_VLDL_CE	703	0.75	0.62	0.904	0.003
Gp	703	1.26	1.08	1.469	0.003
L_LDL_CE	703	0.76	0.63	0.913	0.003
XL_HDL_P	703	1.28	1.08	1.507	0.003
XL_HDL_FC	703	1.27	1.08	1.498	0.003
S_VLDL_C	703	0.75	0.61	0.908	0.004
XL_HDL_L	703	1.28	1.08	1.504	0.004
L_LDL_C	703	0.77	0.64	0.922	0.004
Alb	703	0.77	0.64	0.923	0.005
S_HDL_C	703	0.77	0.64	0.924	0.005
ApoB	703	0.76	0.63	0.926	0.006

D6 continued.

Metabolite	N	HR	Lower CI	Upper CI	*p*-value
		Model 2			
Ace	696	1.30	1.12	1.51	0.000
Crea	685	0.67	0.52	0.86	0.002
FAw3_FA	680	0.73	0.60	0.89	0.002
FAw3	680	0.73	0.59	0.90	0.003
His	703	0.77	0.65	0.92	0.004
		Model 3			
Ace	696	1.30	1.11	1.51	0.001
Crea	685	0.68	0.53	0.89	0.004
		Model 4			
Ace	696	1.28	1.10	1.49	0.002

Abbreviations: ***CI****, 95% confidence interval;* ***HR****, hazard ration;* ***N****, sample number. For metabolite abbreviations, see Appendix D1.*

References:

1. Introduction to Head & Neck Cancer: National Cancer Institute SEER Training Modules; [cited 2018 20.04.2018]. Available from: https://training.seer.cancer.gov/head-neck/intro/ accessed 02.12.2018.
2. Healthcare services for head and neck cancers. Understanding NICE guidance – information for the public: National Institute for Clinical Excellence (NICE); 2017 [21.11.17]. Available from: https://www.nice.org.uk/guidance/csg6/resources/healthcare-services-for-head-and-neck-cancers-pdf-2190221821 accessed 21.11.17.
3. Sanderson RJ, Ironside JA. Squamous cell carcinomas of the head and neck. *BMJ* 2002;325(7368):822-7. [published Online First: 2002/10/12]
4. Heroiu Cataloiu AD, Danciu CE, Popescu CR. Multiple cancers of the head and neck. *Maedica (Buchar)* 2013;8(1):80-5. [published Online First: 2013/09/12]
5. Tshering Vogel DW, Zbaeren P, Thoeny HC. Cancer of the oral cavity and oropharynx. *Cancer Imaging* 2010;10:62-72. doi: 10.1102/1470-7330.2010.0008 [published Online First: 2010/03/18]
6. Anatomy of the Head & Neck: National Cancer Network SEER Training Modules,; [Available from: https://training.seer.cancer.gov/head-neck/anatomy/ accessed 02.12.2018.
7. Head and Neck Cancer Guide: Laryngeal Cancer: Thyroid Head and Neck Cancer Foundation (THANC); 2018 [04.05.2018]. Available from: https://headandneckcancerguide.org/adults/introduction-to-head-and-neck-cancer/throat-cancer/laryngopharyngeal-cancer/laryngeal-cancer/anatomy/ accessed 04.05.2018.
8. Pracy P, Loughran S, Good J, et al. Hypopharyngeal cancer: United Kingdom National Multidisciplinary Guidelines. *The Journal of Laryngology and Ontology* 2016;130(Supp 2):S104-S10. doi: 10.1017/S0022215116000529
9. American Society of Clinical Oncology (ASCO). Nasal Cavity and Paranasal Sinus Cancer: Introduction 2018 [Available from: https://www.cancer.net/cancer-types/nasal-cavity-and-paranasal-sinus-cancer/introduction accessed 20.04.2018.
10. Ibrahim MI, Jusoh YR, Adam NN, et al. Primary Squamous Cell Carcinoma of the Thyroid Gland. *Iran J Otorhinolaryngol* 2018;30(96):65-68. [published Online First: 2018/02/02]
11. Manvikar V, Ramulu S, Ravishanker ST, et al. Squamous cell carcinoma of submandibular salivary gland: A rare case report. *J Oral Maxillofac Pathol* 2014;18(2):299-302. doi: 10.4103/0973-029X.140909 [published Online First: 2014/10/21]
12. World Health Organisation (WHO). Classifications 2018 [Available from: http://www.who.int/classifications/icd/en/ accessed 24.04.2018.
13. Panwar A, Interval E, Lydiatt WM. Emergence of a Novel Staging System for Oropharyngeal Squamous Cell Carcinoma Based on HPV Status. *Oncology (Williston Park)* 2017;31(12):e33-e40. [published Online First: 2018/01/04]
14. World Health Organisation (WHO). International statistical classification of diseases and related health problems 10th Revision Volume 2 Instruction Manual 2016 [Fifth Edition:[Available from: http://apps.who.int/classifications/icd10/browse/Content/statichtml/ICD10Volume2_en_2016.pdf accessed 02.12.2018.
15. Conway DI, Purkayastha M, Chestnutt IG. The changing epidemiology of oral cancer: definitions, trends, and risk factors. *Br Dent J* 2018;225(9):867-73. doi: 10.1038/sj.bdj.2018.922 [published Online First: 2018/11/10]

16. Gillison ML. Human papillomavirus-associated head and neck cancer is a distinct epidemiologic, clinical, and molecular entity. *Semin Oncol* 2004;31(6):744-54. doi: 10.1053/j.seminoncol.2004.09.011 [published Online First: 2004/12/16]
17. Chaturvedi AK, Engels EA, Anderson WF, et al. Incidence trends for human papillomavirus-related and -unrelated oral squamous cell carcinomas in the United States. *J Clin Oncol* 2008;26(4):612-9. doi: 10.1200/JCO.2007.14.1713 [published Online First: 2008/02/01]
18. World Health Organisation. International Statistical Classification of Diseases and Related Health Problems 10th Revision [Available from: https://icd.who.int/browse10/2019/en accessed 23.06.20.
19. MacMillan Cancer Support. Signs And Symptoms Of Head And Neck Cancer 2018 [18.04.2018]. Available from: https://www.macmillan.org.uk/information-and-support/head-and-neck-cancers/understanding-cancer/symptoms.html.
20. Scottish Intercollegiate Guidelines network. Diagnosis and management of head and neck cancer: a national clinical guideline 2006 [Available from: https://www.uhb.nhs.uk/Downloads/pdf/CancerPbDiagnosisHeadAndNeckCancer.pdf .
21. History and Physical Exam: National Cancer Institute SEER Training Modules; [02.12.2018]. Available from: https://training.seer.cancer.gov/diagnostic/history.html accessed 02.12.2018.
22. Hornig JD, Malin BT, O'Connell B. Clinical Evaluation of the Head and Neck Cancer Patient. Head and neck cancer: a multidisciplinary approach. Fourth Electronic Edition ed. Philadelphia: LWW 2014:77-86.
23. The Oral Cancer Foundation. Cancer screening protocols 2018 [19.02.2020]. Available from: https://oralcancerfoundation.org/discovery-diagnosis/cancer-screening-protocols/ accessed 19.02.2020.
24. Cancer Council NSW. Tests for head and neck cancers [19.02.2020]. Available from: https://www.cancercouncil.com.au/head-and-neck-cancer/diagnosis/tests/ accessed 19.02.2020.
25. Mouth Cancer: Diagnosis: NHS; 2016 [Available from: https://www.nhs.uk/conditions/mouth-cancer/diagnosis/ accessed 04.05.2018.
26. Olliff J RP, Conor S, Wong WL, et al. Recommendations for cross-sectional imaging in cancer management. Second ed. London: The Royal College of Radiology, 2014.
27. Wippold FJ. Head and Neck Imaging: The Role of CT and MRI. *Journal of Magnetic Imaging* 2007;25:453-65.
28. Recommendations for cross-sectional imaging in cancer management: Head and Neck cancers: Royal College of Radiologists; 2014 [2nd [accessed 04.05.2018.
29. What is cancer staging? : American Joint Committee on Cancer (AJCC) 2018 [Available from: https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx accessed 21.04.2018.
30. Ferlito A, Shaha AR, Silver CE, et al. Incidence and sites of distant metastases from head and neck cancer. *ORL J Otorhinolaryngol Relat Spec* 2001;63(4):202-7. doi: 10.1159/000055740 [published Online First: 2001/06/16]
31. Jerjes W, Upile T, Radhi H, et al. The effect of tobacco and alcohol and their reduction/cessation on mortality in oral cancer patients: short communication. *Head Neck Oncol* 2012;4:6. doi: 10.1186/1758-3284-4-6
32. Koch WM, Ridge JA, Forastiere A, et al. Comparison of clinical and pathological staging in head and neck squamous cell carcinoma: results from Intergroup Study ECOG 4393/RTOG 9614. *Arch Otolaryngol Head Neck Surg* 2009;135(9):851-8. doi: 10.1001/archoto.2009.123 [published Online First: 2009/09/23]
33. Choi N, Noh Y, Lee EK, et al. Discrepancy between cTNM and pTNM staging of oral cavity cancers and its prognostic significance. *J Surg Oncol* 2017;115(8):1011-18. doi: 10.1002/jso.24606 [published Online First: 2017/03/24]
34. Staging: MacMillan Cancer; 2018 [25.04.2018]. Available from: https://www.macmillan.org.uk/information-and-support/head-and-neck-

cancers/treating/treatment-decisions/understanding-your-diagnosis/staging.html accessed 25.04.2018.
35. Deschler DG, Moore MG, Smith RV. Quick Reference Guide to TNM Staging of Head and Neck Cancer and Neck Dissection Classification, 4th ed.: American Academy of Otolaryngology-Head and Neck Surgery Foundation; 2014 [fourth:[Available from: http://www.entnet.org/sites/default/files/NeckDissection_QuickRefGuide_highresFINAL.pdf accessed 21.04.2018.
36. Staging: National Cancer Institute SEER Training Modules,; [Available from: https://training.seer.cancer.gov/head-neck/abstract-code-stage/staging/ accessed 02.12.2018.
37. American Joint Commitee on Cancer. AJCC Cancer Staging Manual, Eigth Edition.
38. Deschler DG, Day T. Pocket Guide To: TNM Staging of Head and Neck Cancer and Neck Dissection Classification
, Alexandria, VA: American Academy of Otolaryngology– Head and Neck Surgery Foundation, Inc; 2008 [Available from: http://www.sld.cu/galerias/pdf/sitios/cirugiamaxilo/neckdissectionpart1.pdf accessed 30.04.2018.
39. Schroeff MPVd. Prognostic models in Head and Neck Cancer Oncology Predictors and dynamics, 2011.
40. Lydiatt WM, Patel SG, O'Sullivan B, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin* 2017;67(2):122-37. doi: 10.3322/caac.21389 [published Online First: 2017/01/28]
41. Rogers SN, Semple C, Babb M, et al. Quality of life considerations in head and neck cancer: United Kingdom National Multidisciplinary Guidelines. *The Journal of Laryngology and Otology* 2016;130(Supp 2):S49-S52.
42. Head & Neck Cancer Treatment: National Cancer Institute SEER Training Modules; [Available from: https://training.seer.cancer.gov/head-neck/treatment/ accessed 02.12.2018.
43. Healthcare services for head and neck cancers. Understanding NICE guidance – information for the public.: National Institute for Clinical Excellence (NICE),; 2017 [Available from: https://www.nice.org.uk/guidance/csg6/resources/healthcare-services-for-head-and-neck-cancers-pdf-2190221821 accessed 21.11.17.
44. Head and Neck Cancer: Treatment Options: American Society of Clinical Oncology (ASCO); 2018 [Available from: https://www.cancer.net/cancer-types/head-and-neck-cancer/treatment-options accessed 04.05.2018.
45. Head and Neck Cancer: Treatment Options: Cancer.Net; 2017 [Available from: https://www.cancer.net/cancer-types/head-and-neck-cancer/treatment-options accessed 02.12.2018.
46. Bourhis J, Overgaard J, Audry H, et al. Hyperfractionated or accelerated radiotherapy in head and neck cancer: a meta-analysis. *Lancet* 2006;368(9538):843-54. doi: 10.1016/S0140-6736(06)69121-6 [published Online First: 2006/09/05]
47. IMRT and VMAT: The Christie NHS Foundation Trust; [Available from: http://www.christie.nhs.uk/patients-and-visitors/your-treatment-and-care/treatments/radiotherapy/what-we-do/imrt-and-vmat/ accessed 03.05.2018 2018.
48. Nutting CM, Morden JP, Harrington KJ, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. *Lancet Oncol* 2011;12(2):127-36. doi: 10.1016/S1470-2045(10)70290-4 [published Online First: 2011/01/18]
49. Pignon JP, Bourhis J, Domenge C, et al. Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. MACH-NC Collaborative Group. Meta-Analysis of Chemotherapy on Head and Neck Cancer. *Lancet* 2000;355(9208):949-55. [published Online First: 2000/04/18]

50. Cetuximab for treating recurrent or metastatic squamous cell cancer of the head and neck: NICE; 2018 [Available from: https://www.nice.org.uk/guidance/TA473/chapter/1-Recommendations accessed 04.05.2018.
51. Rubin Grandis J, Melhem MF, Gooding WE, et al. Levels of TGF-alpha and EGFR protein in head and neck squamous cell carcinoma and patient survival. *J Natl Cancer Inst* 1998;90(11):824-32. [published Online First: 1998/06/13]
52. Suh Y, Amelio I, Guerrero Urbano T, et al. Clinical update on cancer: molecular oncology of head and neck cancer. *Cell Death Dis* 2014;5:e1018. doi: 10.1038/cddis.2013.548 [published Online First: 2014/01/25]
53. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 2015;136(5):E359-86. doi: 10.1002/ijc.29210 [published Online First: 2014/09/16]
54. Kulkarni MR. Head and neck cancer burden in India. *International Journal of Head and Neck Surgery* 2013;4(1):29-35.
55. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol* 2009;45(4-5):309-16. doi: 10.1016/j.oraloncology.2008.06.002 [published Online First: 2008/09/23]
56. Lakshmaiah KC, Guruprasad B, Lokesh KN, et al. Cancer notification in India. *South Asian J Cancer* 2014;3(1):74-7. doi: 10.4103/2278-330X.126542 [published Online First: 2014/03/26]
57. Reliable monitoring of cancer cases and deaths is essential for successful cancer management plans: American Cancer Society; [11.08.2018]. Available from: http://canceratlas.cancer.org/taking-action/cancer-registries/ accessed 11.08.2018.
58. Mishra A, Meherotra R. Head and neck cancer: global burden and regional trends in India. *Asian Pac J Cancer Prev* 2014;15(2):537-50. [published Online First: 2014/02/27]
59. Swaminathan R, Rama R, Shanta V. Lack of active follow-up of cancer patients in Chennai, India: implications for population-based survival estimates. *Bull World Health Organ* 2008;86(7):509-15. [published Online First: 2008/08/02]
60. Coelho KR. Challenges of the oral cancer burden in India. *J Cancer Epidemiol* 2012;2012:701932. doi: 10.1155/2012/701932 [published Online First: 2012/10/25]
61. The rich picture on people with head and neck cancer: Macmillan 2017 [08.08.2018]. Available from: https://www.macmillan.org.uk/_images/Head-Neck-Cancer_tcm9-282784.pdf accessed 08.08.2018.
62. Ferlay J EM, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F Cancer Today: International Agency for Research into Cancer (IARC), WHO; 2018 [12.09.2018]. Available from: http://gco.iarc.fr/today/home accessed 12.09.2018.
63. Understanding cancer statistics - incidence, survival, mortality: Cancer Research UK (CRUK); 2017 [10.09.2018]. Available from: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/understanding-cancer-statistics-incidence-survival-mortality accessed 10.09.2018.
64. Head and neck cancer incidence statistics: Cancer Research UK (CRUK); [cited 2017 13.09.2018]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers/incidence accessed 13/09/2018.
65. Compendium of Population Health Indicators Annex 3: Health and Social Care Information centre; 2015 [Available from: https://digital.nhs.uk/binaries/content/documents/corporate-website/publication-system/ci-hub/compendium-indicators/compendium-indicators/publicationsystem:cilandingasset%5B3%5D/publicationsystem:Attachments%5B3%5D/publicationsystem:attachmentResource accessed 10.09.2018.
66. Age-standardized Rates: Statistics Canada; 2017 [Available from: https://www.statcan.gc.ca/eng/dai/btd/asr accessed 10.09.2018.

67. Cancer Research UK. Head and Neck Cancer Statistics [08.01.2019]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers#heading-Three accessed 08.01.2019.
68. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68(6):394-424. doi: 10.3322/caac.21492 [published Online First: 2018/09/13]
69. Cancer e-Atlas: ncras; 1999 [14.08.2018]. Available from: http://www.ncin.org.uk/cancer_information_tools/eatlas/ accessed 14.08.2018.
70. McCarthy CE, Field JK, Rajlawat BP, et al. Trends and regional variation in the incidence of head and neck cancers in England: 2002 to 2011. *Int J Oncol* 2015;47(1):204-10. doi: 10.3892/ijo.2015.2990 [published Online First: 2015/05/09]
71. Office for National Statistics. Cancer registration statistics, England [09.01.2019]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancerregistrationstatisticscancerregistrationstatisticsengland accessed 09.01.2019.
72. Adult smoking habits in the UK: 2016: Office for National Statistics (ONS); 2017 [Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2016 accessed 21.11.17.
73. Drinkaware. Consumption: alcohol drinking in the UK 2019 [Available from: https://www.drinkaware.co.uk/research/data/consumption-uk/ accessed 14.02.2019.
74. Fone DL, Farewell DM, White J, et al. Socioeconomic patterning of excess alcohol consumption and binge drinking: a cross-sectional study of multilevel associations with neighbourhood deprivation. *BMJ Open* 2013;3(4) doi: 10.1136/bmjopen-2012-002337 [published Online First: 2013/04/17]
75. Siegler V, Al-Hamad A, Johnson B, et al. Social inequalities in alcohol-related adult mortality by National Statistics Socio-economic Classification, England and Wales, 2001-03. *Health Stat Q* 2011(50):4-39. doi: 10.1057/hsq.2011.7 [published Online First: 2011/06/08]
76. Law MR, Morris JK. Why is mortality higher in poorer areas and in more northern areas of England and Wales? *J Epidemiol Community Health* 1998;52(6):344-52. [published Online First: 1998/10/09]
77. NHS Digital. Statistics on Smoking, England - 2015 2015 [Available from: https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-smoking/statistics-on-smoking-england-2015 accessed 16.01.2019.
78. Beard E, Brown J, West R, et al. Healthier central England or North-South divide? Analysis of national survey data on smoking and high-risk drinking. *BMJ Open* 2017;7(3):e014210. doi: 10.1136/bmjopen-2016-014210 [published Online First: 2017/03/03]
79. Office for National Statistics. Likelihood of smoking four times higher in England's most deprived areas than least deprived 2018 [Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/drugusealcoholandsmoking/articles/likelihoodofsmokingfourtimeshigherinenglandsmostdeprivedareasthanleastdeprived/2018-03-14 accessed 14.02.2019.
80. Office for National Statistics. Do smoking rates vary between more and less advantaged areas? 2014 [Available from: https://webarchive.nationalarchives.gov.uk/20160105204521/http://www.ons.gov.uk/ons/rel/disability-and-health-measurement/do-smoking-rates-vary-between-more-and-less-advantaged-areas-/2012/sty-smoking-rates.html accessed 16.01.2019.
81. Twigg L, Moon G. The spatial and temporal development of binge drinking in England 2001-2009: an observational study. *Soc Sci Med* 2013;91:162-7. doi: 10.1016/j.socscimed.2013.03.023 [published Online First: 2013/04/24]

82. Robinson M, Shipton D, Walsh D, et al. Regional alcohol consumption and alcohol-related mortality in Great Britain: novel insights using retail sales data. *BMC Public Health* 2015;15:1. doi: 10.1186/1471-2458-15-1 [published Online First: 2015/01/08]
83. National Cancer Information Service (NCIS). Head & Neck Cancer Network e-Atlas.
84. Gupta B, Ariyawardana A, Johnson NW. Oral cancer in India continues in epidemic proportions: evidence base and policy initiatives. *Int Dent J* 2013;63(1):12-25. doi: 10.1111/j.1875-595x.2012.00131.x [published Online First: 2013/02/16]
85. World Health Organization. Global Adult Tobacco Survey Fact sheet India 2016-2017 [Available from: https://www.who.int/tobacco/surveillance/survey/gats/GATS_India_2016-17_FactSheet.pdf accessed 14.01.2019.
86. International Agency for Research on Cancer. IARC monographs on carcinogenic risks to humans, Part E,vol. 100. Review of human carcinogens: personal habits and indoor combustions. [Available from: https://monographs.iarc.fr/ENG/Monographs/vol100E/mono100E.pdf accessed 24.01.2018.
87. Sok Ching Cheong PV, Yang Yi-Hsin, Rosnah B Zain, Alexander Ross Kerr, Newell W Johnson. Oral cancer in South East Asia: Current status and future directions. *Translational Research in Oral Oncology* April, 2017
88. Schensul JJ, Nair S, Bilgi S, et al. Availability, accessibility and promotion of smokeless tobacco in a low-income area of Mumbai. *Tob Control* 2013;22(5):324-30. doi: 10.1136/tobaccocontrol-2011-050148 [published Online First: 2012/03/06]
89. Petti S, Masood M, Scully C. The magnitude of tobacco smoking-betel quid chewing-alcohol drinking interaction effect on oral cancer in South-East Asia. A meta-analysis of observational studies. *PLoS One* 2013;8(11):e78999. doi: 10.1371/journal.pone.0078999 [published Online First: 2013/11/22]
90. Organisation WH. Global Information System on Alcohol and Health (GISAH), 2018.
91. Huu Bich T, Thi Quynh Nga P, Ngoc Quang L, et al. Patterns of alcohol consumption in diverse rural populations in the Asian region. *Glob Health Action* 2009;2 doi: 10.3402/gha.v2i0.2017 [published Online First: 2009/12/23]
92. Krishna Rao SV, Mejia G, Roberts-Thomson K, et al. Epidemiology of oral cancer in Asia in the past decade--an update (2000-2012). *Asian Pac J Cancer Prev* 2013;14(10):5567-77. [published Online First: 2013/12/03]
93. Balaram P, Sridhar H, Rajkumar T, et al. Oral cancer in southern India: the influence of smoking, drinking, paan-chewing and oral hygiene. *Int J Cancer* 2002;98(3):440-5. [published Online First: 2002/03/29]
94. World Health organisation. Global status report on alcohol and health 2018 2018 [Available from: http://apps.who.int/iris/bitstream/handle/10665/274603/9789241565639-eng.pdf?ua=1 accessed 18.01.2019.
95. World Health Oranization. Prevalence of tobacco smoking 2019 [Available from: https://www.who.int/gho/tobacco/use/en/ accessed 31.01.2019.
96. American Cancer Society y IaVS. The Tobacco Atlas 2019 [Available from: https://tobaccoatlas.org/ accessed 14.01.2019.
97. Garavello W, Bertuccio P, Levi F, et al. The oral cancer epidemic in central and eastern Europe. *Int J Cancer* 2010;127(1):160-71. doi: 10.1002/ijc.25019 [published Online First: 2009/11/03]
98. World Health Oranization (WHO). Total alcohol per capita (>15 years of age) consumption, in litres of pure alcohol, projected estimates, 2016 2019 [Available from: http://gamapserver.who.int/mapLibrary/Files/Maps/Global_adult_percapita_consumption_2016.png accessed 31.03.2019.
99. World Health Organisation (WHO). Age-standardized prevalence of tobacco smoking among persons aged 15 years and older, 2015 2019 [Available from:

http://gamapserver.who.int/mapLibrary/Files/Maps/Global_Tobacco_use_2015.png accessed 31.03.2019.
100. Marur S, D'Souza G, Westra WH, et al. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol* 2010;11(8):781-9. doi: 10.1016/S1470-2045(10)70017-6
101. de Martel C, Plummer M, Vignat J, et al. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer* 2017;141(4):664-70. doi: 10.1002/ijc.30716 [published Online First: 2017/04/04]
102. Colpani V, Bidinotto AB, Falavigna M, et al. Prevalence of papillomavirus in Brazil: a systematic review protocol. *BMJ Open* 2016;6(11):e011884. doi: 10.1136/bmjopen-2016-011884 [published Online First: 2016/11/25]
103. American Cancer Society. Risk Factors for Laryngeal and Hypopharyngeal Cancers 2019 [Available from: https://www.cancer.org/cancer/laryngeal-and-hypopharyngeal-cancer/causes-risks-prevention/risk-factors.html accessed 16.01.2019.
104. IARC monographs on the evaluation of carcinogenic risks to humans. Tobacco smoke and involuntary smoking 2004 [1-1438]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK316407/pdf/Bookshelf_NBK316407.pdf83.
105. Islami F, Tramacere I, Rota M, et al. Alcohol drinking and laryngeal cancer: overall and dose-risk relation--a systematic review and meta-analysis. *Oral Oncol* 2010;46(11):802-10. doi: 10.1016/j.oraloncology.2010.07.015 [published Online First: 2010/09/14]
106. Altieri A, Garavello W, Bosetti C, et al. Alcohol consumption and risk of laryngeal cancer. *Oral Oncol* 2005;41(10):956-65. doi: 10.1016/j.oraloncology.2005.02.004 [published Online First: 2005/06/02]
107. Talamini R, Bosetti C, La Vecchia C, et al. Combined effect of tobacco and alcohol on laryngeal cancer risk: a case-control study. *Cancer Causes Control* 2002;13(10):957-64. [published Online First: 2003/02/18]
108. Muller F, Wehbe L. Smoking and smoking cessation in Latin America: a review of the current situation and available treatments. *Int J Chron Obstruct Pulmon Dis* 2008;3(2):285-93. [published Online First: 2008/08/09]
109. Perdomo S RG, Brennan P, Forman D, Sierra MS Etiology of head and neck cancer (C01–14, C32) in Central and South America. In: Cancer in Central and South America. . Cancer in Central and South America. Lyon: International Agency for Research on Cancer, 2016.
110. Stefler D, Murphy M, Irdam D, et al. Smoking and Mortality in Eastern Europe: Results From the PrivMort Retrospective Cohort Study of 177 376 Individuals. *Nicotine Tob Res* 2018;20(6):749-54. doi: 10.1093/ntr/ntx122 [published Online First: 2017/06/03]
111. Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65(2):87-108. doi: 10.3322/caac.21262 [published Online First: 2015/02/06]
112. Yu MC, Yuan JM. Epidemiology of nasopharyngeal carcinoma. *Semin Cancer Biol* 2002;12(6):421-9. [published Online First: 2002/11/27]
113. Mahdavifar N, Ghoncheh M, Mohammadian-Hafshejani A, et al. Epidemiology and Inequality in the Incidence and Mortality of Nasopharynx Cancer in Asia. *Osong Public Health Res Perspect* 2016;7(6):360-72. doi: 10.1016/j.phrp.2016.11.002 [published Online First: 2017/01/06]
114. Jia WH, Luo XY, Feng BJ, et al. Traditional Cantonese diet and nasopharyngeal carcinoma risk: a large-scale case-control study in Guangdong, China. *BMC Cancer* 2010;10:446. doi: 10.1186/1471-2407-10-446 [published Online First: 2010/08/24]
115. Barnes L EJ, Reichart P, Sidransky D. Pathology and Genetics of Head and Neck Tumours. WHO Classification of Tumours, 3rd Edition, Volume 9: International Agency for Research on Cancer; [Available from: http://publications.iarc.fr/Book-And-Report-Series/Who-Iarc-Classification-Of-Tumours/Pathology-And-Genetics-Of-Head-And-Neck-Tumours-2005 accessed 09.01.2019.
116. Nawaz I, Moumad K, Martorelli D, et al. Detection of nasopharyngeal carcinoma in Morocco (North Africa) using a multiplex methylation-specific PCR biomarker assay.

Clin Epigenetics 2015;7:89. doi: 10.1186/s13148-015-0119-8 [published Online First: 2015/08/25]
117. Feng BJ, Jalbout M, Ayoub WB, et al. Dietary risk factors for nasopharyngeal carcinoma in Maghrebian countries. *Int J Cancer* 2007;121(7):1550-5. doi: 10.1002/ijc.22813 [published Online First: 2007/06/22]
118. Luukkaa H. Salivary gland cancer in Finland: incidence, histological distribution, outcome and prognostic factors. University of Turku 2010.
119. Aro K. Aspects of the management of salivary gland mucoepidermoid carcinoma in Finland. University of Helsinki, 2012.
120. Boukheris H, Curtis RE, Land CE, et al. Incidence of carcinoma of the major salivary glands according to the WHO classification, 1992 to 2006: a population-based study in the United States. *Cancer Epidemiol Biomarkers Prev* 2009;18(11):2899-906. doi: 10.1158/1055-9965.EPI-09-0638 [published Online First: 2009/10/29]
121. American Cancer Society. What Are the Risk Factors and Potential Causes for Salivary Gland Cancer? 2019 [Available from: https://www.cancer.org/cancer/salivary-gland-cancer/causes-risks-prevention/risk-factors.html accessed 18.01.2019.
122. Sankaranarayanan R, Masuyer E, Swaminathan R, et al. Head and neck cancer: a global perspective on epidemiology and prognosis. *Anticancer Res* 1998;18(6B):4779-86. [published Online First: 1999/01/19]
123. Simard EP, Torre LA, Jemal A. International trends in head and neck cancer incidence rates: differences by country, sex and anatomic site. *Oral Oncol* 2014;50(5):387-403. doi: 10.1016/j.oraloncology.2014.01.016 [published Online First: 2014/02/18]
124. Price G, Roche, M., Crowther, R. Profile of Head and Neck Cancers in England: Incidence, Mortality and Survival National Cancer Intelligence Network,, 2011.
125. Oxford Cancer Intelligence Unit. Profiles of Head and Neck Cancers in England: Incidence, Mortality and Survival 2010 [Available from: http://www.ncin.org.uk/view?rid=69 accessed 30.04.2018.
126. Pampel FC. Cigarette diffusion and sex differences in smoking. *J Health Soc Behav* 2001;42(4):388-404. [published Online First: 2002/02/08]
127. Davy M. Time and generational trends in smoking among men and women in Great Britain, 1972–2004/05. *Health Statistics quarterly* 2006;32:35-43.
128. Institute of Alcohol Studies. Alcohol consumption Factsheet, 2013.
129. British beer and Pub Association (BBPA). UK Alcohol Consumption 1900-2013 [Available from: http://www.beerandpub.com/industry-briefings/alcohol-consumption-data accessed 20/10/2017.
130. World Health Organization. WHO Global Status Report on Alcohol and Health, 2014 2014 [07.09.2018]. Available from: http://www.who.int/substance_abuse/publications/global_alcohol_report/en/ accessed 07.09.2018.
131. British Dental Association. Support Mouth Cancer Action Month 2017: dentists can save lives 2017 [Available from: https://bda.org/news-centre/latest-news-articles/support-mouth-cancer-action-month-2017-dentists-can-save-lives accessed 31.01.2019.
132. Cancer Research UK (CRUK). Mouth cancer rates are increasing, but why? [Available from: https://scienceblog.cancerresearchuk.org/2015/11/13/mouth-cancer-rates-are-increasing-but-why/ accessed 31.03.2019.
133. Curado MP, Hashibe M. Recent changes in the epidemiology of head and neck cancer. *Curr Opin Oncol* 2009;21(3):194-200. doi: 10.1097/CCO.0b013e32832a68ca [published Online First: 2009/04/14]
134. Boyle P, Ferlay J. Cancer incidence and mortality in Europe, 2004. *Ann Oncol* 2005;16(3):481-8. doi: 10.1093/annonc/mdi098 [published Online First: 2005/02/19]
135. Graham H. Smoking prevalence among women in the European community 1950-1990. *Soc Sci Med* 1996;43(2):243-54. [published Online First: 1996/07/01]
136. World Health Organization. Female smoking [Available from: https://www.who.int/tobacco/en/atlas6.pdf accessed 31.01.2019.

137. Chaturvedi AK, Anderson WF, Lortet-Tieulent J, et al. Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *J Clin Oncol* 2013;31(36):4550-9. doi: 10.1200/JCO.2013.50.3870
138. Ryser MD, Rositch A, Gravitt PE. Modeling of US Human Papillomavirus (HPV) Seroprevalence by Age and Sexual Behavior Indicates an Increasing Trend of HPV Infection Following the Sexual Revolution. *J Infect Dis* 2017;216(5):604-11. doi: 10.1093/infdis/jix333 [published Online First: 2017/09/22]
139. National Cancer Institute. Cancer Stat Facts: Laryngeal Cancer [Available from: https://seer.cancer.gov/statfacts/html/laryn.html accessed 23.01.2019.
140. Cancer Research UK. Head and neck cancers mortality statistics [Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers/mortality#ref-4 accessed 01.11.2018.
141. Baxi SS, Pinheiro LC, Patil SM, et al. Causes of death in long-term survivors of head and neck cancer. *Cancer* 2014;120(10):1507-13. doi: 10.1002/cncr.28588
142. World Health Organisation. Risk factors [Available from: http://www.who.int/topics/risk_factors/en/ accessed 11.10.2018.
143. Xue WQ, Qin HD, Ruan HL, et al. Quantitative association of tobacco smoking with the risk of nasopharyngeal carcinoma: a comprehensive meta-analysis of studies conducted between 1979 and 2011. *Am J Epidemiol* 2013;178(3):325-38. doi: 10.1093/aje/kws479 [published Online First: 2013/06/21]
144. Hashibe M, Brennan P, Chuang SC, et al. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Cancer Epidemiol Biomarkers Prev* 2009;18(2):541-50. doi: 10.1158/1055-9965.EPI-08-0347
145. Farshadpour F, Hordijk GJ, Koole R, et al. Non-smoking and non-drinking patients with head and neck squamous cell carcinoma: a distinct population. *Oral Dis* 2007;13(2):239-43. doi: 10.1111/j.1601-0825.2006.01274.x [published Online First: 2007/02/20]
146. Wiseman SM, Swede H, Stoler DL, et al. Squamous cell carcinoma of the head and neck in nonsmokers and nondrinkers: an analysis of clinicopathologic characteristics and treatment outcomes. *Ann Surg Oncol* 2003;10(5):551-7. [published Online First: 2003/06/10]
147. Agudelo D, Quer M, Leon X, et al. Laryngeal carcinoma in patients without a history of tobacco and alcohol use. *Head Neck* 1997;19(3):200-4. [published Online First: 1997/05/01]
148. Shaw R, Beasley N. Aetiology and risk factors for head and neck cancer: United Kingdom National Multidisciplinary Guidelines. *J Laryngol Otol* 2016;130(S2):S9-S12. doi: 10.1017/S0022215116000360 [published Online First: 2016/11/15]
149. Czerninski R, Zini A, Sgan-Cohen HD. Lip cancer: incidence, trends, histology and survival: 1970-2006. *Br J Dermatol* 2010;162(5):1103-9. doi: 10.1111/j.1365-2133.2010.09698.x [published Online First: 2010/02/19]
150. Paget-Bailly S, Cyr D, Luce D. Occupational exposures and cancer of the larynx-systematic review and meta-analysis. *J Occup Environ Med* 2012;54(1):71-84. doi: 10.1097/JOM.0b013e31823c1343 [published Online First: 2011/12/14]
151. Carton M, Barul C, Menvielle G, et al. Occupational exposure to solvents and risk of head and neck cancer in women: a population-based case-control study in France. *BMJ Open* 2017;7(1):e012833. doi: 10.1136/bmjopen-2016-012833 [published Online First: 2017/01/11]
152. Paget-Bailly S, Cyr D, Luce D. Occupational exposures to asbestos, polycyclic aromatic hydrocarbons and solvents, and cancers of the oral cavity and pharynx: a quantitative literature review. *Int Arch Occup Environ Health* 2012;85(4):341-51. doi: 10.1007/s00420-011-0683-y [published Online First: 2011/07/26]
153. WHO Framework Convention on Tobacco Control: World Heath Organisation (WHO); 2005 [03.12.2017]. Available from:

http://apps.who.int/iris/bitstream/10665/42811/1/9241591013.pdf?ua=1 accessed 03.12.2017.
154. Hashibe M, Brennan P, Benhamou S, et al. Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *J Natl Cancer Inst* 2007;99(10):777-89. doi: 10.1093/jnci/djk179
155. Franceschi S, Talamini R, Barra S, et al. Smoking and drinking in relation to cancers of the oral cavity, pharynx, larynx, and esophagus in northern Italy. *Cancer Res* 1990;50(20):6502-7. [published Online First: 1990/10/15]
156. Marron M, Boffetta P, Zhang ZF, et al. Cessation of alcohol drinking, tobacco smoking and the reversal of head and neck cancer risk. *International journal of epidemiology* 2010;39(1):182-96. doi: 10.1093/ije/dyp291
157. Bosetti C, Gallus S, Peto R, et al. Tobacco smoking, smoking cessation, and cumulative risk of upper aerodigestive tract cancers. *Am J Epidemiol* 2008;167(4):468-73. doi: 10.1093/aje/kwm318 [published Online First: 2007/12/07]
158. Bosetti C, Garavello W, Gallus S, et al. Effects of smoking cessation on the risk of laryngeal cancer: an overview of published studies. *Oral Oncol* 2006;42(9):866-72. doi: 10.1016/j.oraloncology.2006.02.008 [published Online First: 2006/08/26]
159. Schlecht NF, Franco EL, Pintos J, et al. Effect of smoking cessation and tobacco type on the risk of cancers of the upper aero-digestive tract in Brazil. *Epidemiology* 1999;10(4):412-8. [published Online First: 1999/07/13]
160. Altieri A, Bosetti C, Talamini R, et al. Cessation of smoking and drinking and the risk of laryngeal cancer. *Br J Cancer* 2002;87(11):1227-9. doi: 10.1038/sj.bjc.6600638 [published Online First: 2002/11/20]
161. Menvielle G, Luce D, Goldberg P, et al. Smoking, alcohol drinking and cancer risk for various sites of the larynx and hypopharynx. A case-control study in France. *Eur J Cancer Prev* 2004;13(3):165-72. [published Online First: 2004/05/29]
162. Hoffmann D, Hoffmann I, El-Bayoumy K. The less harmful cigarette: a controversial issue. a tribute to Ernst L. Wynder. *Chem Res Toxicol* 2001;14(7):767-90. [published Online First: 2001/07/17]
163. Hecht SS. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nat Rev Cancer* 2003;3(10):733-44. doi: 10.1038/nrc1190
164. Pfeifer GP, Denissenko MF, Olivier M, et al. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 2002;21(48):7435-51. doi: 10.1038/sj.onc.1205803 [published Online First: 2002/10/16]
165. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General: U.S. Department of Health and Human Services; 2010 [04.12.2017]. Available from: http://arfreshair.com/wp-content/uploads/2013/01/2010executive_summary.pdf accessed 04.12.2017.
166. Khariwala SS, Ma B, Ruszczak C, et al. High Level of Tobacco Carcinogen-Derived DNA Damage in Oral Cells Is an Independent Predictor of Oral/Head and Neck Cancer Risk in Smokers. *Cancer Prev Res (Phila)* 2017;10(9):507-13. doi: 10.1158/1940-6207.CAPR-17-0140 [published Online First: 2017/07/07]
167. How smoking causes cancer: Cancer Research UK (CRUK); 2016 [03.12.2017]. Available from: http://www.cancerresearchuk.org/causes-of-cancer/smoking-and-cancer/how-smoking-causes-cancer.
168. Lubin JH, Purdue M, Kelsey K, et al. Total exposure and exposure rate effects for alcohol and smoking and risk of head and neck cancer: a pooled analysis of case-control studies. *Am J Epidemiol* 2009;170(8):937-47. doi: 10.1093/aje/kwp222
169. Purdue MP, Hashibe M, Berthiller J, et al. Type of alcoholic beverage and risk of head and neck cancer--a pooled analysis within the INHANCE Consortium. *Am J Epidemiol* 2009;169(2):132-42. doi: 10.1093/aje/kwn306

170. Johansen D, Friis K, Skovenborg E, et al. Food buying habits of people who buy wine or beer: cross sectional study. *BMJ* 2006;332(7540):519-22. doi: 10.1136/bmj.38694.568981.80 [published Online First: 2006/01/24]
171. Tjonneland A, Gronbaek M, Stripp C, et al. Wine intake and diet in a random sample of 48763 Danish men and women. *Am J Clin Nutr* 1999;69(1):49-54. doi: 10.1093/ajcn/69.1.49 [published Online First: 1999/01/30]
172. Klatsky AL, Armstrong MA, Kipp H. Correlates of alcoholic beverage preference: traits of persons who choose wine, liquor or beer. *Br J Addict* 1990;85(10):1279-89. [published Online First: 1990/10/01]
173. Dal Maso L, La Vecchia C, Polesel J, et al. Alcohol drinking outside meals and cancers of the upper aero-digestive tract. *Int J Cancer* 2002;102(4):435-7. doi: 10.1002/ijc.10723 [published Online First: 2002/10/29]
174. Aggarwal BB, Bhardwaj A, Aggarwal RS, et al. Role of resveratrol in prevention and therapy of cancer: preclinical and clinical studies. *Anticancer Res* 2004;24(5A):2783-840. [published Online First: 2004/11/03]
175. Reidy J, McHugh E, Stassen LF. A review of the relationship between alcohol and oral cancer. *Surgeon* 2011;9(5):278-83. doi: 10.1016/j.surge.2011.01.010 [published Online First: 2011/08/17]
176. Seitz HK, Stickel F. Molecular mechanisms of alcohol-mediated carcinogenesis. *Nat Rev Cancer* 2007;7(8):599-612. doi: 10.1038/nrc2191 [published Online First: 2007/07/25]
177. Poschl G, Seitz HK. Alcohol and cancer. *Alcohol Alcohol* 2004;39(3):155-65. [published Online First: 2004/04/15]
178. Lachenmeier DW, Kanteres F, Rehm J. Carcinogenicity of acetaldehyde in alcoholic beverages: risk assessment outside ethanol metabolism. *Addiction* 2009;104(4):533-50. doi: 10.1111/j.1360-0443.2009.02516.x [published Online First: 2009/04/02]
179. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Human Papillomaviruses. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, No. 90 2007 [06.12.2017]. Available from: http://monographs.iarc.fr/ENG/Monographs/vol90/index.php accessed 06.12.2017.
180. Elrefaey S, Massaro MA, Chiocca S, et al. HPV in oropharyngeal cancer: the basics to know in clinical practice. *Acta Otorhinolaryngol Ital* 2014;34(5):299-309.
181. Rettig E, Kiess AP, Fakhry C. The role of sexual behavior in head and neck cancer: implications for prevention and therapy. *Expert Rev Anticancer Ther* 2015;15(1):35-49. doi: 10.1586/14737140.2015.957189 [published Online First: 2014/09/07]
182. Gillison ML, Alemany L, Snijders PJ, et al. Human papillomavirus and diseases of the upper airway: head and neck cancer and respiratory papillomatosis. *Vaccine* 2012;30 Suppl 5:F34-54. doi: 10.1016/j.vaccine.2012.05.070 [published Online First: 2012/12/05]
183. Kreimer AR, Clifford GM, Boyle P, et al. Human papillomavirus types in head and neck squamous cell carcinomas worldwide: a systematic review. *Cancer Epidemiol Biomarkers Prev* 2005;14(2):467-75. doi: 10.1158/1055-9965.EPI-04-0551
184. Franceschi S, Munoz N, Bosch XF, et al. Human papillomavirus and cancers of the upper aerodigestive tract: a review of epidemiological and experimental evidence. *Cancer Epidemiol Biomarkers Prev* 1996;5(7):567-75.
185. D'Souza G, Kreimer AR, Viscidi R, et al. Case-control study of human papillomavirus and oropharyngeal cancer. *N Engl J Med* 2007;356(19):1944-56. doi: 10.1056/NEJMoa065497 [published Online First: 2007/05/15]
186. Riemer AB, Keskin DB, Zhang G, et al. A conserved E7-derived cytotoxic T lymphocyte epitope expressed on human papillomavirus 16-transformed HLA-A2+ epithelial cancers. *J Biol Chem* 2010;285(38):29608-22. doi: 10.1074/jbc.M110.126722 [published Online First: 2010/07/10]
187. Sinha P, Logan HL, Mendenhall WM. Human papillomavirus, smoking, and head and neck cancer. *Am J Otolaryngol* 2012;33(1):130-6. doi: 10.1016/j.amjoto.2011.02.001 [published Online First: 2011/05/07]

188. Schwartz SM, Daling JR, Doody DR, et al. Oral cancer risk in relation to sexual history and evidence of human papillomavirus infection. *J Natl Cancer I* 1998;90(21):1626-36.
189. Shimakage M, Horii K, Tempaku A, et al. Association of Epstein-Barr virus with oral cancers. *Hum Pathol* 2002;33(6):608-14. [published Online First: 2002/08/02]
190. Prabhu SR WD. Evidence of Epstein–Barr Virus Association with Head and Neck Cancers: A Review. *Journal of the Canadian Dental Association* 2016;86(g2)
191. Evans AS. The spectrum of infections with Epstein-Barr virus: a hypothesis. *J Infect Dis* 1971;124(3):330-7. [published Online First: 1971/09/01]
192. Williams H, Crawford DH. Epstein-Barr virus: the impact of scientific advances on clinical practice. *Blood* 2006;107(3):862-9. doi: 10.1182/blood-2005-07-2702 [published Online First: 2005/10/20]
193. Edefonti V, Hashibe M, Ambrogi F, et al. Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. *Ann Oncol* 2012;23(7):1869-80. doi: 10.1093/annonc/mdr548 [published Online First: 2011/11/30]
194. Schwingshackl L, Missbach B, Konig J, et al. Adherence to a Mediterranean diet and risk of diabetes: a systematic review and meta-analysis. *Public Health Nutr* 2015;18(7):1292-9. doi: 10.1017/S1368980014001542 [published Online First: 2014/08/26]
195. Leoncini E, Edefonti V, Hashibe M, et al. Carotenoid intake and head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Eur J Epidemiol* 2016;31(4):369-83. doi: 10.1007/s10654-015-0036-3 [published Online First: 2015/05/02]
196. Kreimer AR, Randi G, Herrero R, et al. Diet and body mass, and oral and oropharyngeal squamous cell carcinomas: analysis from the IARC multinational case-control study. *Int J Cancer* 2006;118(9):2293-7. doi: 10.1002/ijc.21577
197. Boeing H, Dietrich T, Hoffmann K, et al. Intake of fruits and vegetables and risk of cancer of the upper aero-digestive tract: the prospective EPIC-study. *Cancer Causes Control* 2006;17(7):957-69. doi: 10.1007/s10552-006-0036-4 [published Online First: 2006/07/15]
198. Edefonti V, Hashibe M, Parpinel M, et al. Natural vitamin C intake and the risk of head and neck cancer: A pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Int J Cancer* 2015;137(2):448-62. doi: 10.1002/ijc.29388 [published Online First: 2015/01/30]
199. Franceschi S, Dal Maso L, Levi F, et al. Leanness as early marker of cancer of the oral cavity and pharynx. *Ann Oncol* 2001;12(3):331-6. [published Online First: 2001/05/03]
200. Hashibe M, Sankaranarayanan R, Thomas G, et al. Body mass index, tobacco chewing, alcohol drinking and the risk of oral submucous fibrosis in Kerala, India. *Cancer Causes Control* 2002;13(1):55-64. [published Online First: 2002/03/20]
201. Nieto A, Sanchez MJ, Martinez C, et al. Lifetime body mass index and risk of oral cavity and oropharyngeal cancer by smoking and drinking habits. *Br J Cancer* 2003;89(9):1667-71. doi: 10.1038/sj.bjc.6601347
202. Nieto A, Sanchez MJ, Quintana MJ, et al. BMI throughout life, intake of vitamin supplements and oral cancer in Spain. *IARC Sci Publ* 2002;156:259-61. [published Online First: 2002/12/18]
203. D'Avanzo B, La Vecchia C, Talamini R, et al. Anthropometric measures and risk of cancers of the upper digestive and respiratory tract. *Nutr Cancer* 1996;26(2):219-27. doi: 10.1080/01635589609514478 [published Online First: 1996/01/01]
204. Garavello W, Randi G, Bosetti C, et al. Body size and laryngeal cancer risk. *Ann Oncol* 2006;17(9):1459-63. doi: 10.1093/annonc/mdl166 [published Online First: 2006/07/29]
205. Gaudet MM, Olshan AF, Chuang SC, et al. Body mass index and risk of head and neck cancer in a pooled analysis of case-control studies in the International Head and

Neck Cancer Epidemiology (INHANCE) Consortium. *International journal of epidemiology* 2010;39(4):1091-102. doi: 10.1093/ije/dyp380 [published Online First: 2010/02/04]
206. Jacobs DR, Jr., Gottenborg S. Smoking and weight: the Minnesota Lipid Research Clinic. *Am J Public Health* 1981;71(4):391-6. [published Online First: 1981/04/01]
207. John U, Hanke M, Rumpf HJ, et al. Smoking status, cigarettes per day, and their relationship to overweight and obesity among former and current smokers in a national adult general population sample. *Int J Obes (Lond)* 2005;29(10):1289-94. doi: 10.1038/sj.ijo.0803028 [published Online First: 2005/07/06]
208. Istvan JA, Cunningham TW, Garfinkel L. Cigarette smoking and body weight in the Cancer Prevention Study I. *International journal of epidemiology* 1992;21(5):849-53. [published Online First: 1992/10/01]
209. Lubin JH, Gaudet MM, Olshan AF, et al. Body mass index, cigarette smoking, and alcohol consumption and cancers of the oral cavity, pharynx, and larynx: modeling odds ratios in pooled case-control data. *Am J Epidemiol* 2010;171(12):1250-61. doi: 10.1093/aje/kwq088
210. Nicolotti N, Chuang SC, Cadoni G, et al. Recreational physical activity and risk of head and neck cancer: a pooled analysis within the international head and neck cancer epidemiology (INHANCE) Consortium. *Eur J Epidemiol* 2011;26(8):619-28. doi: 10.1007/s10654-011-9612-3 [published Online First: 2011/08/16]
211. Leitzmann MF, Koebnick C, Freedman ND, et al. Physical activity and head and neck cancer risk. *Cancer Causes Control* 2008;19(10):1391-9. doi: 10.1007/s10552-008-9211-0 [published Online First: 2008/08/16]
212. Guha N, Boffetta P, Wunsch Filho V, et al. Oral health and risk of squamous cell carcinoma of the head and neck and esophagus: results of two multicentric case-control studies. *Am J Epidemiol* 2007;166(10):1159-73. doi: 10.1093/aje/kwm193 [published Online First: 2007/09/01]
213. Hashim D, Sartori S, Brennan P, et al. The role of oral hygiene in head and neck cancer: results from International Head and Neck Cancer Epidemiology (INHANCE) consortium. *Ann Oncol* 2016;27(8):1619-25. doi: 10.1093/annonc/mdw224 [published Online First: 2016/05/29]
214. Wang RS, Hu XY, Gu WJ, et al. Tooth loss and risk of head and neck cancer: a meta-analysis. *PLoS One* 2013;8(8):e71122. doi: 10.1371/journal.pone.0071122 [published Online First: 2013/08/31]
215. Tezal M, Sullivan MA, Hyland A, et al. Chronic periodontitis and the incidence of head and neck squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev* 2009;18(9):2406-12. doi: 10.1158/1055-9965.EPI-09-0334 [published Online First: 2009/09/12]
216. Homann N, Tillonen J, Rintamaki H, et al. Poor dental status increases acetaldehyde production from ethanol in saliva: a possible link to increased oral cancer risk among heavy drinkers. *Oral Oncol* 2001;37(2):153-8. [published Online First: 2001/02/13]
217. Shapiro KB, Hotchkiss JH, Roe DA. Quantitative relationship between oral nitrate-reducing activity and the endogenous formation of N-nitrosoamino acids in humans. *Food Chem Toxicol* 1991;29(11):751-5. [published Online First: 1991/11/01]
218. Huang J, Roosaar A, Axell T, et al. A prospective cohort study on poor oral hygiene and pancreatic cancer risk. *Int J Cancer* 2016;138(2):340-7. doi: 10.1002/ijc.29710 [published Online First: 2015/08/04]
219. Muto M, Hitomi Y, Ohtsu A, et al. Acetaldehyde production by non-pathogenic Neisseria in human oral microflora: implications for carcinogenesis in upper aerodigestive tract. *Int J Cancer* 2000;88(3):342-50. [published Online First: 2000/10/31]
220. Bui TC, Markham CM, Ross MW, et al. Examining the association between oral health and oral HPV infection. *Cancer Prev Res (Phila)* 2013;6(9):917-24. doi: 10.1158/1940-6207.CAPR-13-0081 [published Online First: 2013/08/24]
221. Conway DI, Brenner DR, McMahon AD, et al. Estimating and explaining the effect of education and income on head and neck cancer risk: INHANCE consortium pooled

analysis of 31 case-control studies from 27 countries. *Int J Cancer* 2015;136(5):1125-39. doi: 10.1002/ijc.29063 [published Online First: 2014/07/06]
222. Goldgar DE, Easton DF, Cannon-Albright LA, et al. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *J Natl Cancer Inst* 1994;86(21):1600-8. [published Online First: 1994/11/02]
223. Mork J, Moller B, Glattre E. Familial risk in head and neck squamous cell carcinoma diagnosed before the age of 45: a population-based study. *Oral Oncol* 1999;35(4):360-7. [published Online First: 2000/01/25]
224. Li X, Hemminki K. Familial upper aerodigestive tract cancers: incidence trends, familial clustering and subsequent cancers. *Oral Oncol* 2003;39(3):232-9. [published Online First: 2003/03/06]
225. Cancer Institute NSW. Standardised incidence ratio 2015 [Available from: https://www.cancer.nsw.gov.au/glossary/standardised-incidence-ration-(sir) accessed 19.02.2019.
226. Trizna Z, Schantz SP. Hereditary and environmental factors associated with risk and progression of head and neck cancer. *Otolaryngol Clin North Am* 1992;25(5):1089-103. [published Online First: 1992/10/01]
227. Sturgis EM, Wei Q. Genetic susceptibility--molecular epidemiology of head and neck cancer. *Curr Opin Oncol* 2002;14(3):310-7. [published Online First: 2002/05/01]
228. Hiyama T, Yoshihara M, Tanaka S, et al. Genetic polymorphisms and head and neck cancer risk (Review). *Int J Oncol* 2008;32(5):945-73. [published Online First: 2008/04/22]
229. Cadoni G, Boccia S, Petrelli L, et al. A review of genetic epidemiology of head and neck cancer related to polymorphisms in metabolic genes, cell cycle control and alcohol metabolism. *Acta Otorhinolaryngol Ital* 2012;32(1):1-11. [published Online First: 2012/04/14]
230. Hashibe M, McKay JD, Curado MP, et al. Multiple ADH genes are associated with upper aerodigestive cancers. *Nat Genet* 2008;40(6):707-9. doi: 10.1038/ng.151 [published Online First: 2008/05/27]
231. McKay JD, Truong T, Gaborieau V, et al. A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet* 2011;7(3):e1001333. doi: 10.1371/journal.pgen.1001333
232. Lesseur C, Diergaarde B, Olshan AF, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nat Genet* 2016;48(12):1544-50. doi: 10.1038/ng.3685 [published Online First: 2016/11/01]
233. Wei Q, Yu D, Liu M, et al. Genome-wide association study identifies three susceptibility loci for laryngeal squamous cell carcinoma in the Chinese population. *Nat Genet* 2014;46(10):1110-4. doi: 10.1038/ng.3090 [published Online First: 2014/09/10]
234. Lang J, Song X, Cheng J, et al. Association of GSTP1 Ile105Val polymorphism and risk of head and neck cancers: a meta-analysis of 28 case-control studies. *PLoS One* 2012;7(11):e48132. doi: 10.1371/journal.pone.0048132 [published Online First: 2012/11/13]
235. Hashibe M, Brennan P, Strange RC, et al. Meta- and pooled analyses of GSTM1, GSTT1, GSTP1, and CYP1A1 genotypes and risk of head and neck cancer. *Cancer Epidemiol Biomarkers Prev* 2003;12(12):1509-17. [published Online First: 2003/12/25]
236. Zhuo X, Song J, Liao J, et al. Does CYP2E1 RsaI/PstI polymorphism confer head and neck carcinoma susceptibility?: A meta-analysis based on 43 studies. *Medicine (Baltimore)* 2016;95(43):e5156. doi: 10.1097/MD.0000000000005156 [published Online First: 2016/10/28]
237. Liu L, Wu G, Xue F, et al. Functional CYP1A1 genetic variants, alone and in combination with smoking, contribute to development of head and neck cancers. *Eur J Cancer* 2013;49(9):2143-51. doi: 10.1016/j.ejca.2013.01.028 [published Online First: 2013/03/07]

238. Cascorbi I, Brockmoller J, Mrozikiewicz PM, et al. Homozygous rapid arylamine N-acetyltransferase (NAT2) genotype as a susceptibility factor for lung cancer. *Cancer Res* 1996;56(17):3961-6. [published Online First: 1996/09/01]
239. Zheng Y, Li Y, Teng Y, et al. Association of NAT2 phenotype with risk of head and neck carcinoma: A meta-analysis. *Oncol Lett* 2012;3(2):429-34. doi: 10.3892/ol.2011.493 [published Online First: 2012/06/29]
240. Hein DW, Doll MA, Fretland AJ, et al. Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. *Cancer Epidemiol Biomarkers Prev* 2000;9(1):29-42. [published Online First: 2000/02/10]
241. Morita S, Yano M, Tsujinaka T, et al. Genetic polymorphisms of drug-metabolizing enzymes and susceptibility to head-and-neck squamous-cell carcinoma. *Int J Cancer* 1999;80(5):685-8. [published Online First: 1999/02/27]
242. Gonzalez MV, Alvarez V, Pello MF, et al. Genetic polymorphism of N-acetyltransferase-2, glutathione S-transferase-M1, and cytochromes P450IIE1 and P450IID6 in the susceptibility to head and neck cancer. *J Clin Pathol* 1998;51(4):294-8. [published Online First: 1998/07/11]
243. Hahn M, Hagedorn G, Kuhlisch E, et al. Genetic polymorphisms of drug-metabolizing enzymes and susceptibility to oral cavity cancer. *Oral Oncol* 2002;38(5):486-90. [published Online First: 2002/07/12]
244. Katoh T, Kaneko S, Boissy R, et al. A pilot study testing the association between N-acetyltransferases 1 and 2 and risk of oral squamous cell carcinoma in Japanese people. *Carcinogenesis* 1998;19(10):1803-7. [published Online First: 1998/11/07]
245. Varzim G, Monteiro E, Silva R, et al. Polymorphisms of arylamine N-acetyltransferase (NAT1 and NAT2) and larynx cancer susceptibility. *ORL J Otorhinolaryngol Relat Spec* 2002;64(3):206-12. doi: 10.1159/000058026 [published Online First: 2002/05/31]
246. Chen C, Ricks S, Doody DR, et al. N-Acetyltransferase 2 polymorphisms, cigarette smoking and alcohol consumption, and oral squamous cell cancer risk. *Carcinogenesis* 2001;22(12):1993-9. [published Online First: 2001/12/26]
247. Ying XJ, Dong P, Shen B, et al. Possible association of NAT2 polymorphism with laryngeal cancer risk: an evidence-based meta-analysis. *J Cancer Res Clin Oncol* 2011;137(11):1661-7. doi: 10.1007/s00432-011-1045-6 [published Online First: 2011/08/31]
248. Hosagrahara VP, Rettie AE, Hassett C, et al. Functional analysis of human microsomal epoxide hydrolase genetic variants. *Chem Biol Interact* 2004;150(2):149-59. doi: 10.1016/j.cbi.2004.07.004 [published Online First: 2004/11/13]
249. Amador AG, Righi PD, Radpour S, et al. Polymorphisms of xenobiotic metabolizing genes in oropharyngeal carcinoma. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 2002;93(4):440-5. [published Online First: 2002/05/25]
250. Jourenkova-Mironova N, Mitrunen K, Bouchardy C, et al. High-activity microsomal epoxide hydrolase genotypes and the risk of oral, pharynx, and larynx cancers. *Cancer Res* 2000;60(3):534-6. [published Online First: 2000/02/17]
251. Wenghoefer M, Pesch B, Harth V, et al. Association between head and neck cancer and microsomal epoxide hydrolase genotypes. *Arch Toxicol* 2003;77(1):37-41. doi: 10.1007/s00204-002-0414-y [published Online First: 2002/12/20]
252. Li X, Hu Z, Qu X, et al. Putative EPHX1 enzyme activity is related with risk of lung and upper aerodigestive tract cancers: a comprehensive meta-analysis. *PLoS One* 2011;6(3):e14749. doi: 10.1371/journal.pone.0014749 [published Online First: 2011/03/30]
253. Edenberg HJ. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res Health* 2007;30(1):5-13. [published Online First: 2007/08/28]
254. Peters ES, McClean MD, Liu M, et al. The ADH1C polymorphism modifies the risk of squamous cell carcinoma of the head and neck associated with alcohol and tobacco

use. *Cancer Epidemiol Biomarkers Prev* 2005;14(2):476-82. doi: 10.1158/1055-9965.EPI-04-0431
255. Brennan P, Lewis S, Hashibe M, et al. Pooled analysis of alcohol dehydrogenase genotypes and head and neck cancer: a HuGE review. *Am J Epidemiol* 2004;159(1):1-16.
256. Scharer OD. Nucleotide excision repair in eukaryotes. *Cold Spring Harb Perspect Biol* 2013;5(10):a012609. doi: 10.1101/cshperspect.a012609 [published Online First: 2013/10/03]
257. Kietthubthew S, Sriplung H, Au WW. Genetic and environmental interactions on oral cancer in Southern Thailand. *Environ Mol Mutagen* 2001;37(2):111-6. [published Online First: 2001/03/14]
258. Majumder M, Sikdar N, Ghosh S, et al. Polymorphisms at XPD and XRCC1 DNA repair loci and increased risk of oral leukoplakia and cancer among NAT2 slow acetylators. *Int J Cancer* 2007;120(10):2148-56. doi: 10.1002/ijc.22547 [published Online First: 2007/02/10]
259. Gajecka M, Rydzanicz M, Jaskula-Sztul R, et al. Reduced DNA repair capacity in laryngeal cancer subjects. A comparison of phenotypic and genotypic results. *Adv Otorhinolaryngol* 2005;62:25-37. doi: 10.1159/000082460 [published Online First: 2004/12/21]
260. Sturgis EM, Zheng R, Li L, et al. XPD/ERCC2 polymorphisms and risk of head and neck cancer: a case-control analysis. *Carcinogenesis* 2000;21(12):2219-23. [published Online First: 2001/01/03]
261. Ramachandran S, Ramadas K, Hariharan R, et al. Single nucleotide polymorphisms of DNA repair genes XRCC1 and XPD and its molecular mapping in Indian oral cancer. *Oral Oncol* 2006;42(4):350-62. doi: 10.1016/j.oraloncology.2005.08.010 [published Online First: 2005/12/06]
262. Huang WY, Olshan AF, Schwartz SM, et al. Selected genetic polymorphisms in MGMT, XRCC1, XPD, and XRCC3 and risk of head and neck cancer: a pooled analysis. *Cancer Epidemiol Biomarkers Prev* 2005;14(7):1747-53. doi: 10.1158/1055-9965.EPI-05-0162 [published Online First: 2005/07/21]
263. Dhingra V, Verma J, Misra V, et al. Evaluation of Cyclin D1 expression in Head and Neck Squamous Cell Carcinoma. *J Clin Diagn Res* 2017;11(2):EC01-EC04. doi: 10.7860/JCDR/2017/21760.9329 [published Online First: 2017/04/08]
264. Namazie A, Alavi S, Olopade OI, et al. Cyclin D1 amplification and p16(MTS1/CDK4I) deletion correlate with poor prognosis in head and neck tumors. *Laryngoscope* 2002;112(3):472-81. doi: 10.1097/00005537-200203000-00013 [published Online First: 2002/08/01]
265. Deng L, Zhao XR, Pan KF, et al. Cyclin D1 polymorphism and the susceptibility to NPC using DHPLC. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai)* 2002;34(1):16-20. [published Online First: 2002/04/18]
266. Wong YK, Lin SC, Chang CS, et al. Cyclin D1 genotype in areca-associated oral squamous cell carcinoma. *J Oral Pathol Med* 2003;32(5):265-70. [published Online First: 2003/04/16]
267. Nishimoto IN, Pinheiro NA, Rogatto SR, et al. Cyclin D1 gene polymorphism as a risk factor for squamous cell carcinoma of the upper aerodigestive system in non-alcoholics. *Oral Oncol* 2004;40(6):604-10. doi: 10.1016/j.oraloncology.2003.12.009 [published Online First: 2004/04/06]
268. Monteiro E, Varzim G, Pires AM, et al. Cyclin D1 A870G polymorphism and amplification in laryngeal squamous cell carcinoma: implications of tumor localization and tobacco exposure. *Cancer Detect Prev* 2004;28(4):237-43. doi: 10.1016/j.cdp.2004.04.005 [published Online First: 2004/09/08]
269. Holley SL, Matthias C, Jahnke V, et al. Association of cyclin D1 polymorphism with increased susceptibility to oral squamous cell carcinoma. *Oral Oncol* 2005;41(2):156-60. doi: 10.1016/j.oraloncology.2004.08.005 [published Online First: 2005/02/08]

270. Catarino RJ, Breda E, Coelho V, et al. Association of the A870G cyclin D1 gene polymorphism with genetic susceptibility to nasopharyngeal carcinoma. *Head Neck* 2006;28(7):603-8. doi: 10.1002/hed.20377 [published Online First: 2006/05/13]
271. Shin JM, Kamarajan P, Fenno JC, et al. Metabolomics of Head and Neck Cancer: A Mini-Review. *Front Physiol* 2016;7:526. doi: 10.3389/fphys.2016.00526
272. Lu M, Zhan X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA J* 2018;9(1):77-102. doi: 10.1007/s13167-018-0128-8 [published Online First: 2018/03/09]
273. Warr A, Robert C, Hume D, et al. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda)* 2015;5(8):1543-50. doi: 10.1534/g3.115.018564 [published Online First: 2015/07/04]
274. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499(7457):214-18. doi: 10.1038/nature12213 [published Online First: 2013/06/19]
275. Stransky N, Egloff AM, Tward AD, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* 2011;333(6046):1157-60. doi: 10.1126/science.1208130 [published Online First: 2011/07/30]
276. Sepiashvili L, Bruce JP, Huang SH, et al. Novel insights into head and neck cancer using next-generation "omic" technologies. *Cancer Res* 2015;75(3):480-6. doi: 10.1158/0008-5472.CAN-14-3124 [published Online First: 2015/01/16]
277. Perdomo S, Anantharaman D, Foll M, et al. Genomic analysis of head and neck cancer cases from two high incidence regions. *PLoS One* 2018;13(1):e0191701. doi: 10.1371/journal.pone.0191701 [published Online First: 2018/01/30]
278. Beck TN, Golemis EA. Genomic insights into head and neck cancer. *Cancers Head Neck* 2016;1 doi: 10.1186/s41199-016-0003-z [published Online First: 2016/01/01]
279. Castilho RM, Squarize CH, Almeida LO. Epigenetic Modifications and Head and Neck Cancer: Implications for Tumor Progression and Resistance to Therapy. *Int J Mol Sci* 2017;18(7) doi: 10.3390/ijms18071506 [published Online First: 2017/07/15]
280. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 2002;3(6):415-28. doi: 10.1038/nrg816 [published Online First: 2002/06/04]
281. Gasche JA, Goel A. Epigenetic mechanisms in oral carcinogenesis. *Future Oncol* 2012;8(11):1407-25. doi: 10.2217/fon.12.138 [published Online First: 2012/11/15]
282. Mascolo M, Siano M, Ilardi G, et al. Epigenetic disregulation in oral cancer. *Int J Mol Sci* 2012;13(2):2331-53. doi: 10.3390/ijms13022331 [published Online First: 2012/03/13]
283. Zhou C, Ye M, Ni S, et al. DNA methylation biomarkers for head and neck squamous cell carcinoma. *Epigenetics* 2018;13(4):398-409. doi: 10.1080/15592294.2018.1465790 [published Online First: 2018/06/22]
284. Demokan S, Dalay N. Role of DNA methylation in head and neck cancer. *Clin Epigenetics* 2011;2(2):123-50. doi: 10.1007/s13148-011-0045-3 [published Online First: 2012/06/19]
285. Piyathilake CJ, Bell WC, Jones J, et al. Pattern of nonspecific (or global) DNA methylation in oral carcinogenesis. *Head Neck* 2005;27(12):1061-7. doi: 10.1002/hed.20288 [published Online First: 2005/09/13]
286. Irimie AI, Ciocan C, Gulei D, et al. Current Insights into Oral Cancer Epigenetics. *Int J Mol Sci* 2018;19(3) doi: 10.3390/ijms19030670 [published Online First: 2018/03/03]
287. Kudo Y, Kitajima S, Ogawa I, et al. Invasion and metastasis of oral cancer cells require methylation of E-cadherin and/or degradation of membranous beta-catenin. *Clin Cancer Res* 2004;10(16):5455-63. doi: 10.1158/1078-0432.CCR-04-0372 [published Online First: 2004/08/26]
288. Chang HW, Chow V, Lam KY, et al. Loss of E-cadherin expression resulting from promoter hypermethylation in oral tongue carcinoma and its prognostic significance. *Cancer* 2002;94(2):386-92. doi: 10.1002/cncr.10211 [published Online First: 2002/03/20]

289. The GeneCards human gene database. CDKN2A Gene [04.12.19]. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDKN2A&keywords=CDKN2A accessed 04.12.19.
290. The GeneCards human gene database. CDKN2B Gene [04.12.19]. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDKN2B accessed 04.12.19.
291. Gonzalez-Ramirez I, Ramirez-Amador V, Irigoyen-Camacho ME, et al. hMLH1 promoter methylation is an early event in oral cancer. *Oral Oncol* 2011;47(1):22-6. doi: 10.1016/j.oraloncology.2010.10.002 [published Online First: 2010/11/16]
292. The GeneCards human gene database. MLH1 gene [04.12.19]. Available from: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MLH1 accessed 04.12.19.
293. van Kempen PM, Noorlag R, Braunius WW, et al. Differences in methylation profiles between HPV-positive and HPV-negative oropharynx squamous cell carcinoma: a systematic review. *Epigenetics* 2014;9(2):194-203. doi: 10.4161/epi.26881 [published Online First: 2013/10/31]
294. Colacino JA, Dolinoy DC, Duffy SA, et al. Comprehensive analysis of DNA methylation in head and neck squamous cell carcinoma indicates differences by survival and clinicopathologic characteristics. *PLoS One* 2013;8(1):e54742. doi: 10.1371/journal.pone.0054742 [published Online First: 2013/01/30]
295. Koffler J, Sharma S, Hess J. Predictive value of epigenetic alterations in head and neck squamous cell carcinoma. *Mol Cell Oncol* 2014;1(2):e954827. doi: 10.1080/23723548.2014.954827
296. Cancer Genome Atlas N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015;517(7536):576-82. doi: 10.1038/nature14129 [published Online First: 2015/01/30]
297. Seiwert TY, Zuo Z, Keck MK, et al. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. *Clin Cancer Res* 2015;21(3):632-41. doi: 10.1158/1078-0432.CCR-13-3310 [published Online First: 2014/07/25]
298. Agrawal N, Frederick MJ, Pickering CR, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* 2011;333(6046):1154-7. doi: 10.1126/science.1206923 [published Online First: 2011/07/30]
299. Singh NN, Peer A, Nair S, et al. Epigenetics: A possible answer to the undeciphered etiopathogenesis and behavior of oral lesions. *J Oral Maxillofac Pathol* 2016;20(1):122-8. doi: 10.4103/0973-029X.180967 [published Online First: 2016/05/20]
300. Glazer CA, Chang SS, Ha PK, et al. Applying the molecular biology and epigenetics of head and neck cancer in everyday clinical practice. *Oral Oncol* 2009;45(4-5):440-6. doi: 10.1016/j.oraloncology.2008.05.013 [published Online First: 2008/08/05]
301. Shaw RJ, Liloglou T, Rogers SN, et al. Promoter methylation of P16, RARbeta, E-cadherin, cyclin A1 and cytoglobin in oral cancer: quantitative evaluation using pyrosequencing. *Br J Cancer* 2006;94(4):561-8. doi: 10.1038/sj.bjc.6602972 [published Online First: 2006/02/02]
302. Carvalho AL, Jeronimo C, Kim MM, et al. Evaluation of promoter hypermethylation detection in body fluids as a screening/diagnosis tool for head and neck squamous cell carcinoma. *Clin Cancer Res* 2008;14(1):97-107. doi: 10.1158/1078-0432.CCR-07-0722 [published Online First: 2008/01/04]
303. Supic G, Kozomara R, Jovic N, et al. Prognostic significance of tumor-related genes hypermethylation detected in cancer-free surgical margins of oral squamous cell carcinomas. *Oral Oncol* 2011;47(8):702-8. doi: 10.1016/j.oraloncology.2011.05.014 [published Online First: 2011/06/24]
304. Viswanathan M, Tsuchida N, Shanmugam G. Promoter hypermethylation profile of tumor-associated genes p16, p15, hMLH1, MGMT and E-cadherin in oral squamous cell carcinoma. *Int J Cancer* 2003;105(1):41-6. doi: 10.1002/ijc.11028 [published Online First: 2003/04/03]

305. Worsham MJ, Stephen JK, Chen KM, et al. Delineating an epigenetic continuum in head and neck cancer. *Cancer Lett* 2014;342(2):178-84. doi: 10.1016/j.canlet.2012.02.018 [published Online First: 2012/03/06]
306. Ogi K, Toyota M, Ohe-Toyota M, et al. Aberrant methylation of multiple genes and clinicopathological features in oral squamous cell carcinoma. *Clin Cancer Res* 2002;8(10):3164-71. [published Online First: 2002/10/11]
307. Yeh KT, Chang JG, Lin TH, et al. Epigenetic changes of tumor suppressor genes, P15, P16, VHL and P53 in oral cancer. *Oncol Rep* 2003;10(3):659-63. [published Online First: 2003/04/10]
308. Kulkarni V, Saranath D. Concurrent hypermethylation of multiple regulatory genes in chewing tobacco associated oral squamous cell carcinomas and adjacent normal tissues. *Oral Oncol* 2004;40(2):145-53. [published Online First: 2003/12/25]
309. Kaur J, Demokan S, Tripathi SC, et al. Promoter hypermethylation in Indian primary oral squamous cell carcinoma. *Int J Cancer* 2010;127(10):2367-73. doi: 10.1002/ijc.25377 [published Online First: 2010/05/18]
310. Towle R, Truong D, Hogg K, et al. Global analysis of DNA methylation changes during progression of oral cancer. *Oral Oncol* 2013;49(11):1033-42. doi: 10.1016/j.oraloncology.2013.08.005 [published Online First: 2013/09/17]
311. Shiah SG, Chang LC, Tai KY, et al. The involvement of promoter methylation and DNA methyltransferase-1 in the regulation of EpCAM expression in oral squamous cell carcinoma. *Oral Oncol* 2009;45(1):e1-8. doi: 10.1016/j.oraloncology.2008.03.003 [published Online First: 2008/05/20]
312. Chang KW, Kao SY, Tzeng RJ, et al. Multiple molecular alterations of FHIT in betel-associated oral carcinoma. *J Pathol* 2002;196(3):300-6. doi: 10.1002/path.1047 [published Online First: 2002/02/22]
313. Czerninski R, Krichevsky S, Ashhab Y, et al. Promoter hypermethylation of mismatch repair genes, hMLH1 and hMSH2 in oral squamous cell carcinoma. *Oral Dis* 2009;15(3):206-13. doi: 10.1111/j.1601-0825.2008.01510.x [published Online First: 2009/02/12]
314. Gao S, Krogdahl A, Eiberg H, et al. LOH at chromosome 9q34.3 and the Notch1 gene methylation are less involved in oral squamous cell carcinomas. *J Oral Pathol Med* 2007;36(3):173-6. doi: 10.1111/j.1600-0714.2007.00520.x [published Online First: 2007/02/20]
315. Long NK, Kato K, Yamashita T, et al. Hypermethylation of the RECK gene predicts poor prognosis in oral squamous cell carcinomas. *Oral Oncol* 2008;44(11):1052-8. doi: 10.1016/j.oraloncology.2008.02.004 [published Online First: 2008/05/20]
316. Gao F, Huang C, Lin M, et al. Frequent inactivation of RUNX3 by promoter hypermethylation and protein mislocalization in oral squamous cell carcinomas. *J Cancer Res Clin Oncol* 2009;135(5):739-47. doi: 10.1007/s00432-008-0508-x [published Online First: 2008/11/19]
317. Sogabe Y, Suzuki H, Toyota M, et al. Epigenetic inactivation of SFRP genes in oral squamous cell carcinoma. *Int J Oncol* 2008;32(6):1253-61. [published Online First: 2008/05/24]
318. Breitling LP, Yang R, Korn B, et al. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet* 2011;88(4):450-7. doi: 10.1016/j.ajhg.2011.03.003
319. Boscolo-Rizzo P, Furlan C, Lupato V, et al. Novel insights into epigenetic drivers of oropharyngeal squamous cell carcinoma: role of HPV and lifestyle factors. *Clin Epigenetics* 2017;9:124. doi: 10.1186/s13148-017-0424-5 [published Online First: 2017/12/07]
320. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. *Front Genet* 2013;4:132. doi: 10.3389/fgene.2013.00132
321. Pavlova NN, Thompson CB. The Emerging Hallmarks of Cancer Metabolism. *Cell Metab* 2016;23(1):27-47. doi: 10.1016/j.cmet.2015.12.006

322. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013
323. Katada S, Imhof A, Sassone-Corsi P. Connecting threads: epigenetics and metabolism. *Cell* 2012;148(1-2):24-8. doi: 10.1016/j.cell.2012.01.001 [published Online First: 2012/01/24]
324. Sandulache VC, Myers JN. Altered metabolism in head and neck squamous cell carcinoma: an opportunity for identification of novel biomarkers and drug targets. *Head Neck* 2012;34(2):282-90. doi: 10.1002/hed.21664 [published Online First: 2011/02/16]
325. Srivastava S. RR, Gupta V., Tiwari A., Srivastava A. N., Sonkar A. A. Proton HR-MAS MR spectroscopy of oral squamous cell carcinoma tissues: an ex vivo study to identify malignancy induced metabolic fingerprints. *Metabolomics* 2011;7:278–88.
326. Somashekar BS, Kamarajan P, Danciu T, et al. Magic angle spinning NMR-based metabolic profiling of head and neck squamous cell carcinoma tissues. *J Proteome Res* 2011;10(11):5232-41. doi: 10.1021/pr200800w [published Online First: 2011/10/04]
327. Mukherji SK, Schiro S, Castillo M, et al. Proton MR spectroscopy of squamous cell carcinoma of the extracranial head and neck: in vitro and in vivo studies. *AJNR Am J Neuroradiol* 1997;18(6):1057-72. [published Online First: 1997/06/01]
328. Tripathi P, Kamarajan P, Somashekar BS, et al. Delineating metabolic signatures of head and neck squamous cell carcinoma: Phospholipase A(2), a potential therapeutic target. *Int J Biochem Cell B* 2012;44(11):1852-61.
329. Tripathi P, Kamarajan P, Somashekar BS, et al. Delineating metabolic signatures of head and neck squamous cell carcinoma: phospholipase A2, a potential therapeutic target. *Int J Biochem Cell Biol* 2012;44(11):1852-61. doi: 10.1016/j.biocel.2012.06.025 [published Online First: 2012/06/30]
330. Ogawa T, Washio J, Takahashi T, et al. Glucose and glutamine metabolism in oral squamous cell carcinoma: insight from a quantitative metabolomic approach. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2014;118(2):218-25. doi: 10.1016/j.oooo.2014.04.003 [published Online First: 2014/06/15]
331. Sant'Anna-Silva ACB, Santos GC, Campos SPC, et al. Metabolic Profile of Oral Squamous Carcinoma Cell Lines Relies on a Higher Demand of Lipid Metabolism in Metastatic Cells. *Front Oncol* 2018;8:13. doi: 10.3389/fonc.2018.00013 [published Online First: 2018/02/20]
332. Yonezawa K, Nishiumi S, Kitamoto-Matsuda J, et al. Serum and tissue metabolomics of head and neck cancer. *Cancer Genomics Proteomics* 2013;10(5):233-8.
333. El-Sayed S, Bezabeh T, Odlum O, et al. An ex vivo study exploring the diagnostic potential of 1H magnetic resonance spectroscopy in squamous cell carcinoma of the head and neck region. *Head Neck* 2002;24(8):766-72. doi: 10.1002/hed.10125 [published Online First: 2002/08/31]
334. Tiziani S, Lopes V, Gunther UL. Early stage diagnosis of oral cancer using 1H NMR-based metabolomics. *Neoplasia* 2009;11(3):269-76, 4p following 69.
335. Zhou J, Xu B, Huang J, et al. 1H NMR-based metabonomic and pattern recognition analysis for detection of oral squamous cell carcinoma. *Clin Chim Acta* 2009;401(1-2):8-13. doi: 10.1016/j.cca.2008.10.030 [published Online First: 2008/12/06]
336. Almadori G, Bussu F, Galli J, et al. Salivary glutathione and uric acid levels in patients with head and neck squamous cell carcinoma. *Head Neck* 2007;29(7):648-54. doi: 10.1002/hed.20579 [published Online First: 2007/02/03]
337. Yan SK, Wei BJ, Lin ZY, et al. A metabonomic approach to the diagnosis of oral squamous cell carcinoma, oral lichen planus and oral leukoplakia. *Oral Oncol* 2008;44(5):477-83. doi: 10.1016/j.oraloncology.2007.06.007 [published Online First: 2007/10/16]
338. Sugimoto M, Wong DT, Hirayama A, et al. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic

cancer-specific profiles. *Metabolomics* 2010;6(1):78-95. doi: 10.1007/s11306-009-0178-y [published Online First: 2010/03/20]
339. Wei J, Xie G, Zhou Z, et al. Salivary metabolite signatures of oral cancer and leukoplakia. *Int J Cancer* 2011;129(9):2207-17. doi: 10.1002/ijc.25881 [published Online First: 2010/12/31]
340. Xie GX, Chen TL, Qiu YP, et al. Urine metabolite profiling offers potential early diagnosis of oral cancer. *Metabolomics* 2012;8:220–31.
341. Ohshima M, Sugahara K, Kasahara K, et al. Metabolomic analysis of the saliva of Japanese patients with oral squamous cell carcinoma. *Oncol Rep* 2017;37(5):2727-34. doi: 10.3892/or.2017.5561 [published Online First: 2017/04/11]
342. Ishikawa S, Sugimoto M, Kitabatake K, et al. Identification of salivary metabolomic biomarkers for oral cancer screening. *Sci Rep* 2016;6:31520. doi: 10.1038/srep31520 [published Online First: 2016/08/20]
343. Lohavanichbutr P, Zhang Y, Wang P, et al. Salivary metabolite profiling distinguishes patients with oral cavity squamous cell carcinoma from normal controls. *PLoS One* 2018;13(9):e0204249. doi: 10.1371/journal.pone.0204249 [published Online First: 2018/09/21]
344. Vsiansky V, Svobodova M, Gumulec J, et al. Prognostic Significance of Serum Free Amino Acids in Head and Neck Cancers. *Cells* 2019;8(5) doi: 10.3390/cells8050428 [published Online First: 2019/05/12]
345. Mukherjee PK, Funchain P, Retuerto M, et al. Metabolomic analysis identifies differentially produced oral metabolites, including the oncometabolite 2-hydroxyglutarate, in patients with head and neck squamous cell carcinoma. *BBA Clin* 2017;7:8-15. doi: 10.1016/j.bbacli.2016.12.001 [published Online First: 2017/01/06]
346. Cancer Research UK. Head and neck cancer survival statistics [29.01.18]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers/survival#heading-Zero accessed 29.01.18 2018.
347. Fakhry C, Westra WH, Li S, et al. Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. *J Natl Cancer Inst* 2008;100(4):261-9. doi: 10.1093/jnci/djn011
348. O'Rorke MA, Ellison MV, Murray LJ, et al. Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis. *Oral Oncol* 2012;48(12):1191-201. doi: 10.1016/j.oraloncology.2012.06.019
349. Mirghani H, Amen F, Blanchard P, et al. Treatment de-escalation in HPV-positive oropharyngeal carcinoma: ongoing trials, critical issues and perspectives. *Int J Cancer* 2015;136(7):1494-503. doi: 10.1002/ijc.28847 [published Online First: 2014/03/14]
350. Masterson L, Moualed D, Liu ZW, et al. De-escalation treatment protocols for human papillomavirus-associated oropharyngeal squamous cell carcinoma: a systematic review and meta-analysis of current clinical trials. *Eur J Cancer* 2014;50(15):2636-48. doi: 10.1016/j.ejca.2014.07.001 [published Online First: 2014/08/06]
351. Wirth LJ, Burtness B, Nathan CO, et al. Point/Counterpoint: Do We De-escalate Treatment of HPV-Associated Oropharynx Cancer Now? And How? *Am Soc Clin Oncol Educ Book* 2019;39:364-72. doi: 10.1200/EDBK_238315 [published Online First: 2019/05/18]
352. Somlo G. Prognostic Factors (Slides With Transcript) 1999 [Available from: https://www.medscape.org/viewarticle/566795 accessed 07.10.18.
353. Chang TS, Chang CM, Ho HC, et al. Impact of young age on the prognosis for oral cancer: a population-based study in Taiwan. *PLoS One* 2013;8(9):e75855. doi: 10.1371/journal.pone.0075855 [published Online First: 2013/10/03]
354. Vincent N, Dassonville O, Chamorey E, et al. Clinical and histological prognostic factors in locally advanced oral cavity cancers treated with primary surgery. *Eur Ann Otorhinolaryngol Head Neck Dis* 2012;129(6):291-6. doi: 10.1016/j.anorl.2012.01.004 [published Online First: 2012/11/15]

355. Pignon JP, le Maitre A, Maillard E, et al. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17,346 patients. *Radiother Oncol* 2009;92(1):4-14. doi: 10.1016/j.radonc.2009.04.014 [published Online First: 2009/05/19]
356. Sommers LW, Steenbakkers R, Bijl HP, et al. Survival Patterns in Elderly Head and Neck Squamous Cell Carcinoma Patients Treated With Definitive Radiation Therapy. *Int J Radiat Oncol Biol Phys* 2017;98(4):793-801. doi: 10.1016/j.ijrobp.2017.02.214 [published Online First: 2017/05/10]
357. VanderWalde NA, Fleming M, Weiss J, et al. Treatment of older patients with head and neck cancer: a review. *Oncologist* 2013;18(5):568-78. doi: 10.1634/theoncologist.2012-0427 [published Online First: 2013/05/03]
358. Honorato J, Rebelo MS, Dias FL, et al. Gender differences in prognostic factors for oral cancer. *Int J Oral Maxillofac Surg* 2015;44(10):1205-11. doi: 10.1016/j.ijom.2015.04.015 [published Online First: 2015/07/18]
359. Garavello W, Spreafico R, Somigliana E, et al. Prognostic influence of gender in patients with oral tongue cancer. *Otolaryngol Head Neck Surg* 2008;138(6):768-71. doi: 10.1016/j.otohns.2008.02.026 [published Online First: 2008/05/28]
360. Roberts JC, Li G, Reitzel LR, et al. No evidence of sex-related survival disparities among head and neck cancer patients receiving similar multidisciplinary care: a matched-pair analysis. *Clin Cancer Res* 2010;16(20):5019-27. doi: 10.1158/1078-0432.CCR-10-0755 [published Online First: 2010/10/15]
361. de Cassia Braga Ribeiro K, Kowalski LP, Latorre Mdo R. Perioperative complications, comorbidities, and survival in oral or oropharyngeal cancer. *Arch Otolaryngol Head Neck Surg* 2003;129(2):219-28. [published Online First: 2003/02/13]
362. Franco EL, Dib LL, Pinto DS, et al. Race and gender influences on the survival of patients with mouth cancer. *J Clin Epidemiol* 1993;46(1):37-46. [published Online First: 1993/01/01]
363. Faye-Lund H, Abdelnoor M. Prognostic factors of survival in a cohort of head and neck cancer patients in Oslo. *Eur J Cancer B Oral Oncol* 1996;32B(2):83-90. [published Online First: 1996/03/01]
364. Gatta G, Botta L, Sanchez MJ, et al. Prognoses and improvement for head and neck cancers diagnosed in Europe in early 2000s: The EUROCARE-5 population-based study. *Eur J Cancer* 2015;51(15):2130-43. doi: 10.1016/j.ejca.2015.07.043 [published Online First: 2015/10/01]
365. Verbrugge LM. Sex differentials in health. *Public Health Rep* 1982;97(5):417-37. [published Online First: 1982/09/01]
366. Molina MA, Cheung MC, Perez EA, et al. African American and poor patients have a dramatically worse prognosis for head and neck cancer: an examination of 20,915 patients. *Cancer* 2008;113(10):2797-806. doi: 10.1002/cncr.23889 [published Online First: 2008/10/08]
367. Goodwin WJ, Thomas GR, Parker DF, et al. Unequal burden of head and neck cancer in the United States. *Head Neck* 2008;30(3):358-71. doi: 10.1002/hed.20710 [published Online First: 2007/11/01]
368. Murdock JM, Gluckman JL. African-American and white head and neck carcinoma patients in a university medical center setting. Are treatments provided and are outcomes similar or disparate? *Cancer* 2001;91(1 Suppl):279-83. [published Online First: 2001/01/10]
369. Nichols AC, Bhattacharyya N. Racial differences in stage and survival in head and neck squamous cell carcinoma. *Laryngoscope* 2007;117(5):770-5. doi: 10.1097/MLG.0b013e318033c800 [published Online First: 2007/05/03]
370. Shin JY, Truong MT. Racial disparities in laryngeal cancer treatment and outcome: A population-based analysis of 24,069 patients. *Laryngoscope* 2015;125(7):1667-74. doi: 10.1002/lary.25212 [published Online First: 2015/02/20]

371. Megwalu UC, Ma Y. Racial disparities in oropharyngeal cancer survival. *Oral Oncol* 2017;65:33-37. doi: 10.1016/j.oraloncology.2016.12.015 [published Online First: 2017/01/23]
372. Settle K, Posner MR, Schumaker LM, et al. Racial survival disparity in head and neck cancer results from low prevalence of human papillomavirus infection in black oropharyngeal cancer patients. *Cancer Prev Res (Phila)* 2009;2(9):776-81. doi: 10.1158/1940-6207.CAPR-09-0149 [published Online First: 2009/07/31]
373. Zandberg DP, Liu S, Goloubeva O, et al. Oropharyngeal cancer as a driver of racial outcome disparities in squamous cell carcinoma of the head and neck: 10-year experience at the University of Maryland Greenebaum Cancer Center. *Head Neck* 2016;38(4):564-72. doi: 10.1002/hed.23933 [published Online First: 2014/12/10]
374. Worsham MJ, Stephen JK, Lu M, et al. Disparate molecular, histopathology, and clinical factors in head and neck squamous cell carcinoma racial groups. *Otolaryngol Head Neck Surg* 2012;147(2):281-8. doi: 10.1177/0194599812440681 [published Online First: 2012/03/14]
375. Reed AL, Califano J, Cairns P, et al. High frequency of p16 (CDKN2/MTS-1/INK4A) inactivation in head and neck squamous cell carcinoma. *Cancer Res* 1996;56(16):3630-3. [published Online First: 1996/08/15]
376. Lydiatt WM, Davidson BJ, Schantz SP, et al. 9p21 deletion correlates with recurrence in head and neck cancer. *Head Neck* 1998;20(2):113-8. [published Online First: 1998/03/04]
377. Silva TA, Ribeiro FL, Oliveira-Neto HH, et al. Dual role of CCL3/CCR1 in oral squamous cell carcinoma: implications in tumor metastasis and local host defense. *Oncol Rep* 2007;18(5):1107-13. [published Online First: 2007/10/05]
378. Robertson G, Greenlaw N, Steering Group Committee for the Scottish Audit of H, et al. Explaining the effects of socio-economic deprivation on survival in a national prospective cohort study of 1909 patients with head and neck cancers. *Cancer Epidemiol* 2010;34(6):682-8. doi: 10.1016/j.canep.2010.05.009 [published Online First: 2010/06/19]
379. McDonald JT, Johnson-Obaseki S, Hwang E, et al. The relationship between survival and socio-economic status for head and neck cancer in Canada. *J Otolaryngol Head Neck Surg* 2014;43:2. doi: 10.1186/1916-0216-43-2
380. Ragin CC, Langevin SM, Marzouk M, et al. Determinants of head and neck cancer survival by race. *Head Neck* 2011;33(8):1092-8. doi: 10.1002/hed.21584 [published Online First: 2010/10/23]
381. Konski A, Berkey BA, Kian Ang K, et al. Effect of education level on outcome of patients treated on Radiation Therapy Oncology Group Protocol 90-03. *Cancer* 2003;98(7):1497-503. doi: 10.1002/cncr.11661 [published Online First: 2003/09/26]
382. Reitzel LR, Nguyen N, Zafereo ME, et al. Neighborhood deprivation and clinical outcomes among head and neck cancer patients. *Health Place* 2012;18(4):861-8. doi: 10.1016/j.healthplace.2012.03.005
383. Mackillop WJ, Zhang-Salomons J, Groome PA, et al. Socioeconomic status and cancer survival in Ontario. *J Clin Oncol* 1997;15(4):1680-9. doi: 10.1200/JCO.1997.15.4.1680 [published Online First: 1997/04/01]
384. Nutting CM, Robinson M, Birchall M. Survival from laryngeal cancer in England and Wales up to 2001. *Br J Cancer* 2008;99 Suppl 1:S38-9. doi: 10.1038/sj.bjc.6604582
385. Taib BG, Rylands J, Povall S, et al. Protocol: systematic review of the association between socio-economic status and survival in adult head and neck cancer. *Syst Rev* 2017;6(1):151. doi: 10.1186/s13643-017-0545-0 [published Online First: 2017/08/05]
386. National Cancer Intelligence Network. Cancer by Deprivation in England: Public Health England; 2014 [01.11.2018]. Available from: http://www.ncin.org.uk/view?rid=2691 accessed 01.11.2018.
387. Auluck A, Walker BB, Hislop G, et al. Socio-economic deprivation: a significant determinant affecting stage of oral cancer diagnosis and survival. *BMC Cancer* 2016;16:569. doi: 10.1186/s12885-016-2579-4 [published Online First: 2016/08/03]

388. Piccirillo JF. Impact of comorbidity and symptoms on the prognosis of patients with oral carcinoma. *Arch Otolaryngol Head Neck Surg* 2000;126(9):1086-8. [published Online First: 2000/09/09]
389. Piccirillo JF, Costas I. The impact of comorbidity on outcomes. *ORL J Otorhinolaryngol Relat Spec* 2004;66(4):180-5. doi: 10.1159/000079875
390. Skillington SA, Kallogjeri D, Lewis JS, Jr., et al. Prognostic Importance of Comorbidity and the Association Between Comorbidity and p16 in Oropharyngeal Squamous Cell Carcinoma. *JAMA Otolaryngol Head Neck Surg* 2016;142(6):568-75. doi: 10.1001/jamaoto.2016.0347 [published Online First: 2016/04/15]
391. Piccirillo JF, Vlahiotis A. Comorbidity in patients with cancer of the head and neck: prevalence and impact on treatment and prognosis. *Curr Oncol Rep* 2006;8(2):123-9.
392. Ferrier MB, Spuesens EB, Le Cessie S, et al. Comorbidity as a major risk factor for mortality and complications in head and neck surgery. *Arch Otolaryngol Head Neck Surg* 2005;131(1):27-32. doi: 10.1001/archotol.131.1.27 [published Online First: 2005/01/19]
393. Datema FR, Ferrier MB, van der Schroeff MP, et al. Impact of comorbidity on short-term mortality and overall survival of head and neck cancer patients. *Head Neck* 2010;32(6):728-36. doi: 10.1002/hed.21245 [published Online First: 2009/10/15]
394. Boje CR, Dalton SO, Gronborg TK, et al. The impact of comorbidity on outcome in 12 623 Danish head and neck cancer patients: a population based study from the DAHANCA database. *Acta Oncol* 2013;52(2):285-93. doi: 10.3109/0284186X.2012.742964 [published Online First: 2013/01/17]
395. Paleri V, Wight RG, Silver CE, et al. Comorbidity in head and neck cancer: a critical appraisal and recommendations for practice. *Oral Oncol* 2010;46(10):712-9. doi: 10.1016/j.oraloncology.2010.07.008
396. Piccirillo JF, Lacy PD, Basu A, et al. Development of a new head and neck cancer-specific comorbidity index. *Arch Otolaryngol Head Neck Surg* 2002;128(10):1172-9.
397. Coebergh JW, Janssen-Heijnen ML, Razenberg PP. Prevalence of co-morbidity in newly diagnosed patients with cancer: a population-based study. *Crit Rev Oncol Hematol* 1998;27(2):97-100. [published Online First: 1998/05/08]
398. Singh B, Cordeiro PG, Santamaria E, et al. Factors associated with complications in microvascular reconstruction of head and neck defects. *Plast Reconstr Surg* 1999;103(2):403-11. [published Online First: 1999/02/09]
399. Borggreven PA, Kuik DJ, Quak JJ, et al. Comorbid condition as a prognostic factor for complications in major surgery of the oral cavity and oropharynx with microvascular soft tissue reconstruction. *Head Neck* 2003;25(10):808-15. doi: 10.1002/hed.10291 [published Online First: 2003/09/11]
400. Borggreven PA, Kuik DJ, Langendijk JA, et al. Severe comorbidity negatively influences prognosis in patients with oral and oropharyngeal cancer after surgical treatment with microvascular reconstruction. *Oral Oncol* 2005;41(4):358-64. doi: 10.1016/j.oraloncology.2004.08.012 [published Online First: 2005/03/29]
401. Sanabria A, Carvalho AL, Vartanian JG, et al. Factors that influence treatment decision in older patients with resectable head and neck cancer. *Laryngoscope* 2007;117(5):835-40. doi: 10.1097/MLG.0b013e3180337827 [published Online First: 2007/05/03]
402. Piccirillo JF. Importance of comorbidity in head and neck cancer. *Laryngoscope* 2000;110(4):593-602. doi: 10.1097/00005537-200004000-00011
403. Kaplan MH, Feinstein AR. The importance of classifying initial co-morbidity in evaluating the outcome of diabetes mellitus. *J Chronic Dis* 1974;27(7-8):387-404. [published Online First: 1974/09/01]
404. National Cancer Institute. NCI Comorbidity Index Overview 2018 [Available from: https://healthcaredelivery.cancer.gov/seermedicare/considerations/comorbidity.html21.03.2019.

405. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40(5):373-83. [published Online First: 1987/01/01]
406. Hall WH, Ramachandran R, Narayan S, et al. An electronic application for rapidly calculating Charlson comorbidity score. *BMC Cancer* 2004;4:94. doi: 10.1186/1471-2407-4-94 [published Online First: 2004/12/22]
407. Boje CR. Impact of comorbidity on treatment outcome in head and neck squamous cell carcinoma - a systematic review. *Radiother Oncol* 2014;110(1):81-90. doi: 10.1016/j.radonc.2013.07.005 [published Online First: 2013/08/21]
408. Leoncini E, Vukovic V, Cadoni G, et al. Clinical features and prognostic factors in patients with head and neck cancer: Results from a multicentric study. *Cancer Epidemiol* 2015;39(3):367-74. doi: 10.1016/j.canep.2015.02.004 [published Online First: 2015/03/17]
409. Cadoni G, Giraldi L, Petrelli L, et al. Prognostic factors in head and neck cancer: a 10-year retrospective analysis in a single-institution in Italy. *Acta Otorhinolaryngol Ital* 2017;37(6):458-66. doi: 10.14639/0392-100X-1246 [published Online First: 2017/07/01]
410. Le Tourneau C, Velten M, Jung GM, et al. Prognostic indicators for survival in head and neck squamous cell carcinomas: analysis of a series of 621 cases. *Head Neck* 2005;27(9):801-8. doi: 10.1002/hed.20254 [published Online First: 2005/08/09]
411. Brockstein B, Haraf DJ, Rademaker AW, et al. Patterns of failure, prognostic factors and survival in locoregionally advanced head and neck cancer treated with concomitant chemoradiotherapy: a 9-year, 337-patient, multi-institutional experience. *Ann Oncol* 2004;15(8):1179-86. doi: 10.1093/annonc/mdh308
412. Head and Neck Cancers in England. Relative survival by age and stage: Oxford Cancer Intelligence Unit (OCIU); 2011 [Available from: http://www.ncin.org.uk/view?rid=801 accessed 09.11.18.
413. Mehanna H, West CM, Nutting C, et al. Head and neck cancer--Part 2: Treatment and prognostic factors. *BMJ* 2010;341:c4690.
414. Noguti J, De Moura CF, De Jesus GP, et al. Metastasis from oral cancer: an overview. *Cancer Genomics Proteomics* 2012;9(5):329-35. [published Online First: 2012/09/20]
415. O'Sullivan B, Huang SH, Su J, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol* 2016;17(4):440-51. doi: 10.1016/S1470-2045(15)00560-4 [published Online First: 2016/03/05]
416. Jones AS, Roland NJ, Field JK, et al. The level of cervical lymph node metastases: their prognostic relevance and relationship with head and neck squamous carcinoma primary sites. *Clin Otolaryngol Allied Sci* 1994;19(1):63-9. [published Online First: 1994/02/01]
417. Roland N, Porter G, Fish B, et al. Tumour assessment and staging: United Kingdom National Multidisciplinary Guidelines. *J Laryngol Otol* 2016;130(S2):S53-S58. doi: 10.1017/S002221511600044X [published Online First: 2016/11/15]
418. Ang KK, Harris J, Wheeler R, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med* 2010;363(1):24-35. doi: 10.1056/NEJMoa0912217
419. Poeta ML, Manola J, Goldwasser MA, et al. TP53 mutations and survival in squamous-cell carcinoma of the head and neck. *N Engl J Med* 2007;357(25):2552-61. doi: 10.1056/NEJMoa073770 [published Online First: 2007/12/21]
420. Lindenbergh-van der Plas M, Brakenhoff RH, Kuik DJ, et al. Prognostic significance of truncating TP53 mutations in head and neck squamous cell carcinoma. *Clin Cancer Res* 2011;17(11):3733-41. doi: 10.1158/1078-0432.CCR-11-0183 [published Online First: 2011/04/07]

421. Omura G, Ando M, Ebihara Y, et al. The prognostic value of TP53 mutations in hypopharyngeal squamous cell carcinoma. *BMC Cancer* 2017;17(1):898. doi: 10.1186/s12885-017-3913-1 [published Online First: 2017/12/29]
422. Cutilli T, Leocata P, Dolo V, et al. p53 as a prognostic marker associated with the risk of mortality for oral squamous cell carcinoma. *Oncol Lett* 2016;12(2):1046-50. doi: 10.3892/ol.2016.4742 [published Online First: 2016/07/23]
423. Tandon S, Tudur-Smith C, Riley RD, et al. A systematic review of p53 as a prognostic factor of survival in squamous cell carcinoma of the four main anatomical subsites of the head and neck. *Cancer Epidemiol Biomarkers Prev* 2010;19(2):574-87. doi: 10.1158/1055-9965.EPI-09-0981 [published Online First: 2010/02/10]
424. Zhao YY, Yu GT, Xiao T, et al. The Notch signaling pathway in head and neck squamous cell carcinoma: A meta-analysis. *Adv Clin Exp Med* 2017;26(5):881-87. doi: 10.17219/acem/64000 [published Online First: 2017/10/27]
425. Fukusumi T, Califano JA. The NOTCH Pathway in Head and Neck Squamous Cell Carcinoma. *J Dent Res* 2018;97(6):645-53. doi: 10.1177/0022034518760297 [published Online First: 2018/03/01]
426. Mriouah J, Boura C, Gargouri M, et al. PTEN expression is involved in the invasive properties of HNSCC: a key protein to consider in locoregional recurrence. *Int J Oncol* 2014;44(3):709-16. doi: 10.3892/ijo.2013.2219 [published Online First: 2013/12/25]
427. Kostareli E, Holzinger D, Bogatyrova O, et al. HPV-related methylation signature predicts survival in oropharyngeal squamous cell carcinomas. *J Clin Invest* 2013;123(6):2488-501. doi: 10.1172/JCI67010
428. Sinha P, Bahadur S, Thakar A, et al. Significance of promoter hypermethylation of p16 gene for margin assessment in carcinoma tongue. *Head Neck* 2009;31(11):1423-30. doi: 10.1002/hed.21122 [published Online First: 2009/05/12]
429. Xing XB, Cai WB, Luo L, et al. The Prognostic Value of p16 Hypermethylation in Cancer: A Meta-Analysis. *PLoS One* 2013;8(6):e66587. doi: 10.1371/journal.pone.0066587 [published Online First: 2013/06/28]
430. Tham T, Bardash Y, Herman SW, et al. Neutrophil-to-lymphocyte ratio as a prognostic indicator in head and neck cancer: A systematic review and meta-analysis. *Head Neck* 2018 doi: 10.1002/hed.25324 [published Online First: 2018/05/16]
431. Templeton AJ, McNamara MG, Seruga B, et al. Prognostic role of neutrophil-to-lymphocyte ratio in solid tumors: a systematic review and meta-analysis. *J Natl Cancer Inst* 2014;106(6):dju124. doi: 10.1093/jnci/dju124 [published Online First: 2014/05/31]
432. Yu Y, Wang H, Yan A, et al. Pretreatment neutrophil to lymphocyte ratio in determining the prognosis of head and neck cancer: a meta-analysis. *BMC Cancer* 2018;18(1):383. doi: 10.1186/s12885-018-4230-z [published Online First: 2018/04/06]
433. Rosculet N, Zhou XC, Ha P, et al. Neutrophil-to-lymphocyte ratio: Prognostic indicator for head and neck squamous cell carcinoma. *Head Neck* 2017;39(4):662-67. doi: 10.1002/hed.24658 [published Online First: 2017/01/12]
434. Cho JK, Kim MW, Choi IS, et al. Optimal cutoff of pretreatment neutrophil-to-lymphocyte ratio in head and neck cancer patients: a meta-analysis and validation study. *BMC Cancer* 2018;18(1):969. doi: 10.1186/s12885-018-4876-6 [published Online First: 2018/10/13]
435. Tham T, Olson C, Khaymovich J, et al. The lymphocyte-to-monocyte ratio as a prognostic indicator in head and neck cancer: a systematic review and meta-analysis. *Eur Arch Otorhinolaryngol* 2018;275(7):1663-70. doi: 10.1007/s00405-018-4972-x [published Online First: 2018/04/14]
436. Kano S, Homma A, Hatakeyama H, et al. Pretreatment lymphocyte-to-monocyte ratio as an independent prognostic factor for head and neck cancer. *Head Neck* 2017;39(2):247-53. doi: 10.1002/hed.24576 [published Online First: 2016/09/13]

437. Takenaka Y, Oya R, Kitamiura T, et al. Platelet count and platelet-lymphocyte ratio as prognostic markers for head and neck squamous cell carcinoma: A meta-analysis. *Head Neck* 2018 doi: 10.1002/hed.25366 [published Online First: 2018/08/14]
438. Pandya PH, Murray ME, Pollok KE, et al. The Immune System in Cancer Pathogenesis: Potential Therapeutic Approaches. *J Immunol Res* 2016;2016:4273943. doi: 10.1155/2016/4273943 [published Online First: 2017/01/25]
439. Silva P, Homer JJ, Slevin NJ, et al. Clinical and biological factors affecting response to radiotherapy in patients with head and neck cancer: a review. *Clin Otolaryngol* 2007;32(5):337-45. doi: 10.1111/j.1749-4486.2007.01544.x [published Online First: 2007/09/22]
440. Gong L, Zhang W, Zhou J, et al. Prognostic value of HIFs expression in head and neck cancer: a systematic review. *PLoS One* 2013;8(9):e75094. doi: 10.1371/journal.pone.0075094 [published Online First: 2013/09/24]
441. Duffy SA, Ronis DL, McLean S, et al. Pretreatment health behaviors predict survival among patients with head and neck squamous cell carcinoma. *J Clin Oncol* 2009;27(12):1969-75. doi: 10.1200/JCO.2008.18.2188
442. Mayne ST, Cartmel B, Kirsh V, et al. Alcohol and tobacco use prediagnosis and postdiagnosis, and survival in a cohort of patients with early stage cancers of the oral cavity, pharynx, and larynx. *Cancer Epidemiol Biomarkers Prev* 2009;18(12):3368-74. doi: 10.1158/1055-9965.epi-09-0944
443. Sharp L, McDevitt J, Carsin AE, et al. Smoking at diagnosis is an independent prognostic factor for cancer-specific survival in head and neck cancer: findings from a large, population-based study. *Cancer Epidemiol Biomarkers Prev* 2014;23(11):2579-90. doi: 10.1158/1055-9965.epi-14-0311
444. Hatcher JL, Sterba KR, Tooze JA, et al. Tobacco Use and Surgical Outcomes in Head and Neck Cancer Patients. *Head Neck* 2014 doi: 10.1002/hed.23944
445. Warren GW, Kasza KA, Reid ME, et al. Smoking at diagnosis and survival in cancer patients. *Int J Cancer* 2013;132(2):401-10. doi: 10.1002/ijc.27617
446. Gillison ML, Zhang Q, Jordan R, et al. Tobacco smoking and increased risk of death and progression for patients with p16-positive and p16-negative oropharyngeal cancer. *J Clin Oncol* 2012;30(17):2102-11. doi: 10.1200/JCO.2011.38.4099 [published Online First: 2012/05/09]
447. Dikshit RP, Boffetta P, Bouchardy C, et al. Lifestyle habits as prognostic factors in survival of laryngeal and hypopharyngeal cancer: a multicentric European study. *Int J Cancer* 2005;117(6):992-5. doi: 10.1002/ijc.21244
448. Dikshit RP, Boffetta P, Bouchardy C, et al. Risk factors for the development of second primary tumors among men after laryngeal and hypopharyngeal carcinoma. *Cancer* 2005;103(11):2326-33. doi: 10.1002/cncr.21051
449. Crosignani P, Russo A, Tagliabue G, et al. Tobacco and diet as determinants of survival in male laryngeal cancer patients. *Int J Cancer* 1996;65(3):308-13. doi: 10.1002/(SICI)1097-0215(19960126)65:3<308::AID-IJC5>3.0.CO;2-3
450. Fortin A, Wang CS, Vigneault E. Influence of smoking and alcohol drinking behaviors on treatment outcomes of patients with squamous cell carcinomas of the head and neck. *Int J Radiat Oncol Biol Phys* 2009;74(4):1062-9. doi: 10.1016/j.ijrobp.2008.09.021
451. Do KA, Johnson MM, Doherty DA, et al. Second primary tumors in patients with upper aerodigestive tract cancers: joint effects of smoking and alcohol (United States). *Cancer Causes Control* 2003;14(2):131-8.
452. Day GL, Blot WJ, Shore RE, et al. Second cancers following oral and pharyngeal cancers: role of tobacco and alcohol. *J Natl Cancer Inst* 1994;86(2):131-7.
453. Saunders JB AO, Babor TF, et al. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption–II. *Addiction* 1993;88:791-804.

454. Lang S, Schimansky S, Beynon R, et al. Dietary behaviors and survival in people with head and neck cancer: Results from Head and Neck 5000. *Head Neck* 2019;41(7):2074-84. doi: 10.1002/hed.25660 [published Online First: 2019/01/31]
455. Boffetta P, Merletti F, Faggiano F, et al. Prognostic factors and survival of laryngeal cancer patients from Turin, Italy. A population-based study. *Am J Epidemiol* 1997;145(12):1100-5.
456. Hafkamp HC, Manni JJ, Haesevoets A, et al. Marked differences in survival rate between smokers and nonsmokers with HPV 16-associated tonsillar carcinomas. *Int J Cancer* 2008;122(12):2656-64. doi: 10.1002/ijc.23458
457. Maxwell JH, Kumar B, Feng FY, et al. Tobacco use in human papillomavirus-positive advanced oropharynx cancer patients related to increased risk of distant metastases and tumor recurrence. *Clin Cancer Res* 2010;16(4):1226-35. doi: 10.1158/1078-0432.CCR-09-2350
458. McBride SM, Ali NN, Margalit DN, et al. Active tobacco smoking and distant metastasis in patients with oropharyngeal cancer. *Int J Radiat Oncol Biol Phys* 2012;84(1):183-8. doi: 10.1016/j.ijrobp.2011.11.044
459. Peterson LA, Bellile EL, Wolf GT, et al. Cigarette use, comorbidities, and prognosis in a prospective head and neck squamous cell carcinoma population. *Head Neck* 2016 doi: 10.1002/hed.24515
460. Degner LF, Kristjanson LJ, Bowman D, et al. Information needs and decisional preferences in women with breast cancer. *JAMA* 1997;277(18):1485-92. [published Online First: 1997/05/14]
461. Davison BJ, Goldenberg SL, Gleave ME, et al. Provision of individualized information to men and their partners to facilitate treatment decision making in prostate cancer. *Oncol Nurs Forum* 2003;30(1):107-14. doi: 10.1188/03.ONF.107-114 [published Online First: 2003/01/08]
462. Steinhauser KE, Christakis NA, Clipp EC, et al. Preparing for the end of life: preferences of patients, families, physicians, and other care providers. *J Pain Symptom Manage* 2001;22(3):727-37. [published Online First: 2001/09/05]
463. Datema FR, Moya A, Krause P, et al. Novel head and neck cancer survival analysis approach: random survival forests versus Cox proportional hazards regression. *Head Neck* 2012;34(1):50-8. doi: 10.1002/hed.21698 [published Online First: 2011/02/16]
464. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10(2):e1001381. doi: 10.1371/journal.pmed.1001381 [published Online First: 2013/02/09]
465. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-87. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 [published Online First: 1996/02/28]
466. Royston P, Moons KG, Altman DG, et al. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;338:b604. doi: 10.1136/bmj.b604 [published Online First: 2009/04/02]
467. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605. doi: 10.1136/bmj.b605 [published Online First: 2009/05/30]
468. Moons KG, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606. doi: 10.1136/bmj.b606 [published Online First: 2009/06/09]
469. Mes SW. tBD, , Poli T., Rossi S., Scheckenbach K., van Wieringen WN., Brink A., Bertani, N., Lanfranco D., Silini EM., van Diest PJ., Bloemena E., Leemans CR., van de Wiel MA., and Brakenhoff RH. Prognostic modeling of oral cancer by gene profiles and clinicopathological co-variables. *Oncotargret* 2017;8(35):59312-23.

470. Almangush A, Coletta RD, Bello IO, et al. A simple novel prognostic model for early stage oral tongue cancer. *Int J Oral Maxillofac Surg* 2015;44(2):143-50. doi: 10.1016/j.ijom.2014.10.004 [published Online First: 2014/12/03]
471. Emerick KS, Leavitt ER, Michaelson JS, et al. Initial clinical findings of a mathematical model to predict survival of head and neck cancer. *Otolaryngol Head Neck Surg* 2013;149(4):572-8. doi: 10.1177/0194599813495178 [published Online First: 2013/06/26]
472. Rios Velazquez E, Hoebers F, Aerts HJ, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. *Radiother Oncol* 2014;113(3):324-30. doi: 10.1016/j.radonc.2014.09.005 [published Online First: 2014/12/03]
473. Baatenburg de Jong RJ, Hermans J, Molenaar J, et al. Prediction of survival in patients with head and neck cancer. *Head Neck* 2001;23(9):718-24.
474. Cao W, Liu JN, Liu Z, et al. A three-lncRNA signature derived from the Atlas of ncRNA in cancer (TANRIC) database predicts the survival of patients with head and neck squamous cell carcinoma. *Oral Oncol* 2017;65:94-101. doi: 10.1016/j.oraloncology.2016.12.017 [published Online First: 2017/01/23]
475. Chen F, Cao Y, Huang J, et al. A novel prognostic index for oral squamous cell carcinoma patients with surgically treated. *Oncotarget* 2017;8(33):55525-33. doi: 10.18632/oncotarget.14821 [published Online First: 2017/09/15]
476. Datema FR, Ferrier MB, van der Schroeff MP, et al. Impact of Comorbidity on Short-Term Mortality and Overall Survival of Head and Neck Cancer Patients. *Head Neck-J Sci Spec* 2010;32(6):728-36.
477. Datema FR, Ferrier MB, Vergouwe Y, et al. Update and external validation of a head and neck cancer prognostic model. *Head Neck* 2013;35(9):1232-7. doi: 10.1002/hed.23117 [published Online First: 2012/08/01]
478. Fakhry C, Zhang Q, Nguyen-Tan PF, et al. Development and Validation of Nomograms Predictive of Overall and Progression-Free Survival in Patients With Oropharyngeal Cancer. *J Clin Oncol* 2017;35(36):4057-65. doi: 10.1200/JCO.2016.72.0748 [published Online First: 2017/08/05]
479. Montero PH, Yu C, Palmer FL, et al. Nomograms for preoperative prediction of prognosis in patients with oral cavity squamous cell carcinoma. *Cancer* 2014;120(2):214-21. doi: 10.1002/cncr.28407 [published Online First: 2014/01/09]
480. Pugliano FA, Piccirillo JF, Zequeira MR, et al. Symptoms as an index of biologic behavior in head and neck cancer. *Otolaryngol Head Neck Surg* 1999;120(3):380-6. doi: 10.1016/S0194-5998(99)70279-2 [published Online First: 1999/03/04]
481. Te Riele R, Dronkers EAC, Wieringa MH, et al. Influence of anemia and BMI on prognosis of laryngeal squamous cell carcinoma: Development of an updated prognostic model. *Oral Oncol* 2018;78:25-30. doi: 10.1016/j.oraloncology.2018.01.001 [published Online First: 2018/03/03]
482. Rietbergen MM, Brakenhoff RH, Bloemena E, et al. Human papillomavirus detection and comorbidity: critical issues in selection of patients with oropharyngeal cancer for treatment De-escalation trials. *Ann Oncol* 2013;24(11):2740-5. doi: 10.1093/annonc/mdt319 [published Online First: 2013/08/16]
483. Rietbergen MM, Witte BI, Velazquez ER, et al. Different prognostic models for different patient populations: validation of a new prognostic model for patients with oropharyngeal cancer in Western Europe. *Br J Cancer* 2015;112(11):1733-6. doi: 10.1038/bjc.2015.139 [published Online First: 2015/05/08]
484. Tertipis N, Hammar U, Nasman A, et al. A model for predicting clinical outcome in patients with human papillomavirus-positive tonsillar and base of tongue cancer. *Eur J Cancer* 2015;51(12):1580-7. doi: 10.1016/j.ejca.2015.04.024 [published Online First: 2015/05/31]
485. Lee T.A. PAS. Exposure Definition and Measurement. In: Velentgas P DN, Nourjah P, et al., ed. Developing a Protocol for Observational Comparative Effectiveness

Research: A User's Guide. Rockville (MD): Agency for Healthcare Research and Quality (US) 2013.
486. Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J Public Health (Oxf)* 2005;27(3):281-91. doi: 10.1093/pubmed/fdi031 [published Online First: 2005/05/05]
487. Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. *Int J Behav Healthc Res* 2011;2(4):320-32. doi: 10.1504/IJBHR.2011.043414 [published Online First: 2011/10/01]
488. Yeager DS, Krosnick JA. The validity of self-reported nicotine product use in the 2001-2008 National Health and Nutrition Examination Survey. *Med Care* 2010;48(12):1128-32. doi: 10.1097/MLR.0b013e3181ef9948 [published Online First: 2010/10/14]
489. Bjerregaard P, Becker U. Validation of survey information on smoking and alcohol consumption against import statistics, Greenland 1993-2010. *Int J Circumpolar Health* 2013;72 doi: 10.3402/ijch.v72i0.20314 [published Online First: 2013/03/09]
490. Poikolainen K, Podkletnova I, Alho H. Accuracy of quantity-frequency and graduated frequency questionnaires in measuring alcohol intake: comparison with daily diary and commonly used laboratory markers. *Alcohol Alcohol* 2002;37(6):573-6. doi: 10.1093/alcalc/37.6.573 [published Online First: 2002/11/05]
491. Strasser AA, Pickworth WB, Patterson F, et al. Smoking topography predicts abstinence following treatment with nicotine replacement therapy. *Cancer Epidemiol Biomarkers Prev* 2004;13(11 Pt 1):1800-4. [published Online First: 2004/11/10]
492. Del Boca FK, Darkes J. The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction* 2003;98 Suppl 2:1-12. [published Online First: 2004/02/27]
493. Graham K, Demers A, Rehm J, et al. Problems with the graduated frequency approach to measuring alcohol consumption: results from a pilot study in Toronto, Canada. *Alcohol Alcohol* 2004;39(5):455-62. doi: 10.1093/alcalc/agh075 [published Online First: 2004/08/04]
494. National Institute on Alcohol Abuse and Alcoholism. Methodological Issues in Measuring Alcohol Use 2003 [30.05.2019]. Available from: https://pubs.niaaa.nih.gov/publications/arh27-1/18-29.htm accessed 30.05.2017.
495. Gmel G, Graham K, Kuendig H, et al. Measuring alcohol consumption--should the 'graduated frequency' approach become the norm in survey research? *Addiction* 2006;101(1):16-30. doi: 10.1111/j.1360-0443.2005.01224.x [published Online First: 2006/01/06]
496. The National Institute on Alcohol Abuse and Alcoholism. Biomarkers of haevy drinking 2004 [Available from: https://pubs.niaaa.nih.gov/publications/assessingalcohol/biomarkers.htm accessed 31.05.2019.
497. Lester BM, Conradt E, Marsit C. Introduction to the Special Section on Epigenetics. *Child Dev* 2016;87(1):29-37. doi: 10.1111/cdev.12489 [published Online First: 2016/01/30]
498. Deans C, Maggert KA. What do you mean, "epigenetic"? *Genetics* 2015;199(4):887-96. doi: 10.1534/genetics.114.173492 [published Online First: 2015/04/10]
499. Guida F, Sandanger TM, Castagne R, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet* 2015;24(8):2349-59. doi: 10.1093/hmg/ddu751 [published Online First: 2015/01/04]
500. Relton CL, Hartwig FP, Davey Smith G. From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive. *International journal of epidemiology* 2015;44(4):1083-93. doi: 10.1093/ije/dyv198 [published Online First: 2015/10/02]
501. Costello JF, Plass C. Methylation matters. *J Med Genet* 2001;38(5):285-303. doi: 10.1136/jmg.38.5.285 [published Online First: 2001/05/23]

502. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 2019;20(10):590-607. doi: 10.1038/s41580-019-0159-6 [published Online First: 2019/08/11]
503. Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 2010;11(3):191-203. doi: 10.1038/nrg2732
504. Barros-Silva D, Marques CJ, Henrique R, et al. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. *Genes (Basel)* 2018;9(9) doi: 10.3390/genes9090429 [published Online First: 2018/08/26]
505. Cabezas J., Lucey M.R. , Bataller R. Biomarkers for Monitoring Alcohol Use. *Clinical liver disease* 2016;8(3)
506. Peterson K. Biomarkers for Alcohol Use and Abuse. *Alcohol Resrearch and Health* 2004/2005;28(1):30-37.
507. Philibert R, Hollenbeck N, Andersen E, et al. A quantitative epigenetic approach for the assessment of cigarette consumption. *Front Psychol* 2015;6:656. doi: 10.3389/fpsyg.2015.00656 [published Online First: 2015/06/18]
508. Philibert RA, Penaluna B, White T, et al. A pilot examination of the genome-wide DNA methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs. *Epigenetics* 2014;9(9):1212-9. doi: 10.4161/epi.32252 [published Online First: 2014/08/26]
509. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet* 2016;9(5):436-47. doi: 10.1161/CIRCGENETICS.116.001506
510. Zhang Y, Schottker B, Florath I, et al. Smoking-Associated DNA Methylation Biomarkers and Their Predictive Value for All-Cause and Cardiovascular Mortality. *Environ Health Perspect* 2016;124(1):67-74. doi: 10.1289/ehp.1409020 [published Online First: 2015/05/29]
511. Liu C, Marioni RE, Hedman AK, et al. A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry* 2016 doi: 10.1038/mp.2016.192
512. Hattab MW, Clark SL, van den Oord E. Overestimation of the classification accuracy of a biomarker for assessing heavy alcohol use. *Mol Psychiatry* 2018;23(11):2114-15. doi: 10.1038/mp.2017.181 [published Online First: 2017/09/13]
513. Karlsson Linner R, Marioni RE, Rietveld CA, et al. An epigenome-wide association study meta-analysis of educational attainment. *Mol Psychiatry* 2017;22(12):1680-90. doi: 10.1038/mp.2017.210 [published Online First: 2017/11/01]
514. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* 2018;19(6):371-84. doi: 10.1038/s41576-018-0004-3 [published Online First: 2018/04/13]
515. Cole JH, Marioni RE, Harris SE, et al. Brain age and other bodily 'ages': implications for neuropsychiatry. *Mol Psychiatry* 2019;24(2):266-81. doi: 10.1038/s41380-018-0098-1 [published Online First: 2018/06/13]
516. Jylhava J, Pedersen NL, Hagg S. Biological Age Predictors. *EBioMedicine* 2017;21:29-36. doi: 10.1016/j.ebiom.2017.03.046 [published Online First: 2017/04/12]
517. Perna L, Zhang Y, Mons U, et al. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. *Clin Epigenetics* 2016;8:64. doi: 10.1186/s13148-016-0228-z [published Online First: 2016/06/09]
518. Fransquet PD, Wrigglesworth J, Woods RL, et al. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. *Clin Epigenetics* 2019;11(1):62. doi: 10.1186/s13148-019-0656-7 [published Online First: 2019/04/13]
519. Bocklandt S, Lin W, Sehl ME, et al. Epigenetic predictor of age. *PLoS One* 2011;6(6):e14821. doi: 10.1371/journal.pone.0014821
520. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 2013;49(2):359-67. doi: 10.1016/j.molcel.2012.10.016
521. Marioni RE, Shah S, McRae AF, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol* 2015;16:25. doi: 10.1186/s13059-015-0584-6

522. Chen BH, Marioni RE, Colicino E, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)* 2016;8(9):1844-65. doi: 10.18632/aging.101020 [published Online First: 2016/10/01]
523. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;14(10):R115. doi: 10.1186/gb-2013-14-10-r115
524. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* 2018;10(4):573-91. doi: 10.18632/aging.101414 [published Online First: 2018/04/21]
525. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* 2019;11(2):303-27. doi: 10.18632/aging.101684 [published Online First: 2019/01/23]
526. Christiansen L, Lenart A, Tan Q, et al. DNA methylation age is associated with mortality in a longitudinal Danish twin study. *Aging Cell* 2016;15(1):149-54. doi: 10.1111/acel.12421
527. Dugue PA, Bassett JK, Joo JE, et al. Association of DNA Methylation-Based Biological Age With Health Risk Factors and Overall and Cause-Specific Mortality. *Am J Epidemiol* 2018;187(3):529-38. doi: 10.1093/aje/kwx291 [published Online First: 2017/10/12]
528. McCartney DL, Stevenson AJ, Walker RM, et al. Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimers Dement (Amst)* 2018;10:429-37. doi: 10.1016/j.dadm.2018.05.006 [published Online First: 2018/09/01]
529. Quach A, Levine ME, Tanaka T, et al. Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY)* 2017;9(2):419-46. doi: 10.18632/aging.101168 [published Online First: 2017/02/16]
530. Scheffner M, Huibregtse JM, Vierstra RD, et al. The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell* 1993;75(3):495-505. [published Online First: 1993/11/05]
531. Alunni-Fabbroni M, Littlewood T, Deleu L, et al. Induction of S phase and apoptosis by the human papillomavirus type 16 E7 protein are separable events in immortalized rodent fibroblasts. *Oncogene* 2000;19(19):2277-85. doi: 10.1038/sj.onc.1203570 [published Online First: 2000/05/24]
532. Chung CH, Gillison ML. Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications. *Clin Cancer Res* 2009;15(22):6758-62. doi: 10.1158/1078-0432.CCR-09-0784 [published Online First: 2009/10/29]
533. Gillison ML, Lowy DR. A causal role for human papillomavirus in head and neck cancer. *Lancet* 2004;363(9420):1488-9. doi: 10.1016/S0140-6736(04)16194-1
534. Qureishi A, Mawby T, Fraser L, et al. Current and future techniques for human papilloma virus (HPV) testing in oropharyngeal squamous cell carcinoma. *Eur Arch Otorhinolaryngol* 2017;274(7):2675-83. doi: 10.1007/s00405-017-4503-1 [published Online First: 2017/03/13]
535. Broglie MA, Jochum W, Forbs D, et al. Brush cytology for the detection of high-risk HPV infection in oropharyngeal squamous cell carcinoma. *Cancer Cytopathol* 2015;123(12):732-8. doi: 10.1002/cncy.21606 [published Online First: 2015/09/09]
536. Qureishi A, Ali M, Fraser L, et al. Saliva testing for human papilloma virus in oropharyngeal squamous cell carcinoma: A diagnostic accuracy study. *Clin Otolaryngol* 2018;43(1):151-57. doi: 10.1111/coa.12917 [published Online First: 2017/06/18]
537. Miller DL, Puricelli MD, Stack MS. Virology and molecular pathogenesis of HPV (human papillomavirus)-associated oropharyngeal squamous cell carcinoma. *Biochem J* 2012;443(2):339-53. doi: 10.1042/BJ20112017 [published Online First: 2012/03/29]
538. Walline HM, Komarck C, McHugh JB, et al. High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and oral cavity cancers: comparison of multiple methods. *JAMA Otolaryngol Head Neck Surg* 2013;139(12):1320-7. doi: 10.1001/jamaoto.2013.5460 [published Online First: 2013/11/02]

539. Allen CT, Lewis JS, Jr., El-Mofty SK, et al. Human papillomavirus and oropharynx cancer: biology, detection and clinical implications. *Laryngoscope* 2010;120(9):1756-72. doi: 10.1002/lary.20936 [published Online First: 2010/07/30]
540. Bishop JA, Ma XJ, Wang H, et al. Detection of transcriptionally active high-risk HPV in patients with head and neck squamous cell carcinoma as visualized by a novel E6/E7 mRNA in situ hybridization method. *Am J Surg Pathol* 2012;36(12):1874-82. doi: 10.1097/PAS.0b013e318265fb2b [published Online First: 2012/10/13]
541. Rooper LM, Gandhi M, Bishop JA, et al. RNA in-situ hybridization is a practical and effective method for determining HPV status of oropharyngeal squamous cell carcinoma including discordant cases that are p16 positive by immunohistochemistry but HPV negative by DNA in-situ hybridization. *Oral Oncol* 2016;55:11-6. doi: 10.1016/j.oraloncology.2016.02.008
542. Wang F, Flanagan J, Su N, et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J Mol Diagn* 2012;14(1):22-9. doi: 10.1016/j.jmoldx.2011.08.002 [published Online First: 2011/12/15]
543. Kim KY, Lewis JS, Jr., Chen Z. Current status of clinical testing for human papillomavirus in oropharyngeal squamous cell carcinoma. *J Pathol Clin Res* 2018;4(4):213-26. doi: 10.1002/cjp2.111 [published Online First: 2018/07/31]
544. Curt Malloy JS, Cristina Herdman. HPV DNA Testing: Technical and Programmatic Issues for Cervical Cancer Prevention in Low-Resource Settings 2000 [Available from: http://screening.iarc.fr/doc/HPV-DNA-Testing-Issues.pdf accessed 13.03.2019.
545. Chernock RD, El-Mofty SK, Thorstad WL, et al. HPV-related nonkeratinizing squamous cell carcinoma of the oropharynx: utility of microscopic features in predicting patient outcome. *Head Neck Pathol* 2009;3(3):186-94. doi: 10.1007/s12105-009-0126-1 [published Online First: 2010/07/03]
546. Dahlstrom KR, Anderson KS, Cheng JN, et al. HPV Serum Antibodies as Predictors of Survival and Disease Progression in Patients with HPV-Positive Squamous Cell Carcinoma of the Oropharynx. *Clin Cancer Res* 2015;21(12):2861-9. doi: 10.1158/1078-0432.CCR-14-3323
547. Kreimer AR, Johansson M, Waterboer T, et al. Evaluation of human papillomavirus antibodies and risk of subsequent head and neck cancer. *J Clin Oncol* 2013;31(21):2708-15. doi: 10.1200/JCO.2012.47.2738
548. Kreimer AR, Johansson M, Yanik EL, et al. Kinetics of the Human Papillomavirus Type 16 E6 Antibody Response Prior to Oropharyngeal Cancer. *J Natl Cancer Inst* 2017;109(8) doi: 10.1093/jnci/djx005 [published Online First: 2017/04/05]
549. Zhang Y, Waterboer T, Haddad RI, et al. Human papillomavirus (HPV) 16 antibodies at diagnosis of HPV-related oropharyngeal cancer and antibody trajectories after treatment. *Oral Oncol* 2017;67:77-82. doi: 10.1016/j.oraloncology.2017.02.004 [published Online First: 2017/03/30]
550. Fakhry C, Qualliotine JR, Zhang Z, et al. Serum Antibodies to HPV16 Early Proteins Warrant Investigation as Potential Biomarkers for Risk Stratification and Recurrence of HPV-Associated Oropharyngeal Cancer. *Cancer Prev Res (Phila)* 2016;9(2):135-41. doi: 10.1158/1940-6207.CAPR-15-0299 [published Online First: 2015/12/25]
551. Rubenstein LM, Smith EM, Pawlita M, et al. Human papillomavirus serologic follow-up response and relationship to survival in head and neck cancer: a case-comparison study. *Infect Agent Cancer* 2011;6:9. doi: 10.1186/1750-9378-6-9 [published Online First: 2011/07/12]
552. Koslabova E, Hamsikova E, Salakova M, et al. Markers of HPV infection and survival in patients with head and neck tumors. *Int J Cancer* 2013;133(8):1832-9. doi: 10.1002/ijc.28194 [published Online First: 2013/04/09]
553. Waterboer T, Sehr P, Pawlita M. Suppression of non-specific binding in serological Luminex assays. *J Immunol Methods* 2006;309(1-2):200-4. doi: 10.1016/j.jim.2005.11.008 [published Online First: 2006/01/13]

554. Waterboer T, Sehr P, Michael KM, et al. Multiplex human papillomavirus serology based on in situ-purified glutathione s-transferase fusion proteins. *Clin Chem* 2005;51(10):1845-53. doi: 10.1373/clinchem.2005.052381
555. Robbins HA, Waterboer T, Porras C, et al. Immunogenicity assessment of HPV16/18 vaccine using the glutathione S-transferase L1 multiplex serology assay. *Hum Vaccin Immunother* 2014;10(10):2965-74. doi: 10.4161/21645515.2014.972811 [published Online First: 2014/12/09]
556. Cell Biolabs I. HPV16 E7 Oncoprotein ELISA Kit 2016 [Available from: https://www.cellbiolabs.com/sites/default/files/VPK-5045-hpv16-e7-elisa-kit.pdf accessed 15.03.2019.
557. Racaniello V. Detection of antigens or antibodies by ELISA 2010 [Available from: http://www.virology.ws/2010/07/16/detection-of-antigens-or-antibodies-by-elisa/ accessed 15.03.2019.
558. abcam. ELISA principle 2019 [Available from: https://www.abcam.com/kits/elisa-principle accessed 15.03.2019.
559. Venuti A, Paolini F. HPV detection methods in head and neck cancer. *Head Neck Pathol* 2012;6 Suppl 1:S63-74. doi: 10.1007/s12105-012-0372-5
560. Schache AG, Liloglou T, Risk JM, et al. Evaluation of human papilloma virus diagnostic testing in oropharyngeal squamous cell carcinoma: sensitivity, specificity, and prognostic discrimination. *Clin Cancer Res* 2011;17(19):6262-71. doi: 10.1158/1078-0432.CCR-11-0388 [published Online First: 2011/10/05]
561. Smeets SJ, Hesselink AT, Speel EJ, et al. A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. *Int J Cancer* 2007;121(11):2465-72. doi: 10.1002/ijc.22980 [published Online First: 2007/08/08]
562. Thavaraj S, Stokes A, Guerra E, et al. Evaluation of human papillomavirus testing for squamous cell carcinoma of the tonsil in clinical practice. *J Clin Pathol* 2011;64(4):308-12. doi: 10.1136/jcp.2010.088450 [published Online First: 2011/02/25]
563. Turi KN, Romick-Rosendale L, Ryckman KK, et al. A review of metabolomics approaches and their application in identifying causal pathways of childhood asthma. *J Allergy Clin Immunol* 2018;141(4):1191-201. doi: 10.1016/j.jaci.2017.04.021 [published Online First: 2017/05/10]
564. Cheng S, Shah SH, Corwin EJ, et al. Potential Impact and Study Considerations of Metabolomics in Cardiovascular Health and Disease: A Scientific Statement From the American Heart Association. *Circ Cardiovasc Genet* 2017;10(2) doi: 10.1161/HCG.0000000000000032 [published Online First: 2017/04/01]
565. Gorrochategui E, Jaumot, J., Lacorte, S., Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends in Analytical Chemistry* 2016;82:425-42.
566. Roberts LD, Souza AL, Gerszten RE, et al. Targeted metabolomics. *Curr Protoc Mol Biol* 2012;Chapter 30:Unit 30 2 1-24. doi: 10.1002/0471142727.mb3002s98 [published Online First: 2012/04/04]
567. Veenstra TD. Metabolomics: the final frontier? *Genome Med* 2012;4(4):40. doi: 10.1186/gm339 [published Online First: 2012/05/02]
568. James TL. Fundamentals of NMR 1998 [03/07/2019]. Available from: https://qudev.phys.ethz.ch/phys4/studentspresentations/nmr/James_Fundamentals_of_NMR.pdf accessed 03.07.2019.
569. Finehout EJ, Lee KH. An introduction to mass spectrometry applications in biological research. *Biochem Mol Biol Educ* 2004;32(2):93-100. doi: 10.1002/bmb.2004.494032020331 [published Online First: 2004/03/01]
570. Emwas AH. The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. *Methods Mol Biol* 2015;1277:161-93. doi: 10.1007/978-1-4939-2377-9_13 [published Online First: 2015/02/14]

571. Emwas AH, Roy R, McKay RT, et al. NMR Spectroscopy for Metabolomics Research. *Metabolites* 2019;9(7) doi: 10.3390/metabo9070123 [published Online First: 2019/06/30]
572. Rattray NJW, Deziel NC, Wallach JD, et al. Beyond genomics: understanding exposotypes through metabolomics. *Hum Genomics* 2018;12(1):4. doi: 10.1186/s40246-018-0134-x [published Online First: 2018/01/28]
573. Ness AR, Waylen A, Hurley K, et al. Recruitment, response rates and characteristics of 5511 people enrolled in a prospective clinical cohort study: head and neck 5000. *Clin Otolaryngol* 2016;41(6):804-09. doi: 10.1111/coa.12548
574. Ness AR, Waylen A, Hurley K, et al. Establishing a large prospective clinical cohort in people with head and neck cancer as a biomedical resource: head and neck 5000. *Bmc Cancer* 2014;14
575. Beynon RA, Lang S, Schimansky S, et al. Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer. *Int J Cancer* 2018;143(5):1114-27. doi: 10.1002/ijc.31416 [published Online First: 2018/04/03]
576. Head and neck 5000. Head and neck 5000 - Number of people with complete baseline data [Available from: http://www.headandneck5000.org.uk/content/5776810f886983.82723170.pdf accessed 21.05.2019.
577. Adult Comorbidity Evaluation-27: NHS; 2018 [Available from: https://www.datadictionary.nhs.uk/data_dictionary/nhs_business_definitions/a/adult_comorbidity_evaluation_-_27_de.asp?shownav=1.
578. Piccirillo JF, Tierney RM, Costas I, et al. Prognostic importance of comorbidity in a hospital-based cancer registry. *JAMA* 2004;291(20):2441-7. doi: 10.1001/jama.291.20.2441 [published Online First: 2004/05/27]
579. Kallogjeri D, Gaynor SM, Piccirillo ML, et al. Comparison of comorbidity collection methods. *J Am Coll Surg* 2014;219(2):245-55. doi: 10.1016/j.jamcollsurg.2014.01.059 [published Online First: 2014/06/17]
580. Head and Neck 5000. Participating centres [Available from: http://www.headandneck5000.org.uk/about-us/participating-centres/ accessed 21.05.2019.
581. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* 2017;26(1):126-35. doi: 10.1158/1055-9965.EPI-16-0106 [published Online First: 2016/10/05]
582. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48(10):1284-87. doi: 10.1038/ng.3656 [published Online First: 2016/08/30]
583. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2011;9(2):179-81. doi: 10.1038/nmeth.1785 [published Online First: 2011/12/06]
584. Howie B, Fuchsberger C, Stephens M, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;44(8):955-9. doi: 10.1038/ng.2354 [published Online First: 2012/07/24]
585. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48(10):1279-83. doi: 10.1038/ng.3643 [published Online First: 2016/08/23]
586. Infinium® OncoArray-500K BeadChips: Illumina; 2014 [Available from: https://support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_oncoarray500k.pdf.
587. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.

Genome Biol 2016;17(1):208. doi: 10.1186/s13059-016-1066-1 [published Online First: 2016/10/09]
588. Min JL, Hemani G, Davey Smith G, et al. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* 2018;34(23):3983-89. doi: 10.1093/bioinformatics/bty476 [published Online First: 2018/06/23]
589. Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010;11:587. doi: 10.1186/1471-2105-11-587 [published Online First: 2010/12/02]
590. Soininen P, Kangas AJ, Wurtz P, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst* 2009;134(9):1781-5. doi: 10.1039/b910205a [published Online First: 2009/08/18]
591. Soininen P, Kangas AJ, Wurtz P, et al. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* 2015;8(1):192-206. doi: 10.1161/CIRCGENETICS.114.000216
592. Inouye M, Kettunen J, Soininen P, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol* 2010;6:441. doi: 10.1038/msb.2010.93
593. Wurtz P, Havulinna AS, Soininen P, et al. Metabolite profiling and cardiovascular event risk: a prospective study of 3 population-based cohorts. *Circulation* 2015;131(9):774-85. doi: 10.1161/CIRCULATIONAHA.114.013116
594. Wurtz P, Kangas AJ, Soininen P, et al. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *Am J Epidemiol* 2017;186(9):1084-96. doi: 10.1093/aje/kwx016
595. Fischer K, Kettunen J, Wurtz P, et al. Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observational study of 17,345 persons. *PLoS Med* 2014;11(2):e1001606. doi: 10.1371/journal.pmed.1001606
596. Wurtz P, Wang Q, Soininen P, et al. Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *Journal of the American College of Cardiology* 2016;67(10):1200-10. doi: 10.1016/j.jacc.2015.12.060
597. Kettunen J, Demirkan A, Wurtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* 2016;7:11122. doi: 10.1038/ncomms11122 [published Online First: 2016/03/24]
598. Wang Q, Wurtz P, Auro K, et al. Effects of hormonal contraception on systemic metabolism: cross-sectional and longitudinal evidence. *International journal of epidemiology* 2016;45(5):1445-57. doi: 10.1093/ije/dyw147
599. Wang Q, Wurtz P, Auro K, et al. Metabolic profiling of pregnancy: cross-sectional and longitudinal evidence. *BMC medicine* 2016;14(1):205. doi: 10.1186/s12916-016-0733-0
600. Wang Q, Kangas AJ, Soininen P, et al. Sex hormone-binding globulin associations with circulating lipids and metabolites and the risk for type 2 diabetes: observational and causal effect estimates. *International journal of epidemiology* 2015;44(2):623-37. doi: 10.1093/ije/dyv093
601. Wurtz P, Wang Q, Kangas AJ, et al. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med* 2014;11(12):e1001765. doi: 10.1371/journal.pmed.1001765
602. Kujala UM, Makinen VP, Heinonen I, et al. Long-term leisure-time physical activity and serum metabolome. *Circulation* 2013;127(3):340-8. doi: 10.1161/CIRCULATIONAHA.112.105551
603. Holmes MV, Millwood IY, Kartsonaki C, et al. Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke. *Journal of the American College of Cardiology* 2018;71(6):620-32. doi: 10.1016/j.jacc.2017.12.006
604. Publications: Nightingale Health; 2018 [23.03.2018]. Available from: https://nightingalehealth.com/publications.
605. Holzinger D, Wichmann G, Baboci L, et al. Sensitivity and specificity of antibodies against HPV16 E6 and other early proteins for the detection of HPV16-driven

oropharyngeal squamous cell carcinoma. *Int J Cancer* 2017;140(12):2748-57. doi: 10.1002/ijc.30697 [published Online First: 2017/03/21]
606. Schmitt M, Bravo IG, Snijders PJ, et al. Bead-based multiplex genotyping of human papillomaviruses. *J Clin Microbiol* 2006;44(2):504-12. doi: 10.1128/JCM.44.2.504-512.2006
607. NHS. Alcohol units 2018 [Available from: https://www.nhs.uk/live-well/alcohol-support/calculating-alcohol-units/19.06.2019.
608. Szklo M. Population-based cohort studies. *Epidemiol Rev* 1998;20(1):81-90. doi: 10.1093/oxfordjournals.epirev.a017974 [published Online First: 1998/10/08]
609. Sawabe M, Ito H, Oze I, et al. Heterogeneous impact of alcohol consumption according to treatment method on survival in head and neck cancer: A prospective study. *Cancer Sci* 2017;108(1):91-100. doi: 10.1111/cas.13115 [published Online First: 2016/11/02]
610. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393. doi: 10.1136/bmj.b2393
611. Shah AD, Bartlett JW, Carpenter J, et al. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol* 2014;179(6):764-74. doi: 10.1093/aje/kwt312 [published Online First: 2014/03/05]
612. Tilling K, Williamson EJ, Spratt M, et al. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *J Clin Epidemiol* 2016;80:107-15. doi: 10.1016/j.jclinepi.2016.07.004 [published Online First: 2016/07/23]
613. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *Stata Journal* 2009;9:466–77.
614. Rubin D, Schenker N Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association* 1989;81:366–74.
615. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009;28(15):1982-98. doi: 10.1002/sim.3618
616. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23(5):723-48. doi: 10.1002/sim.1621 [published Online First: 2004/02/26]
617. Schoenfeld D. Partial Residuals for The Proportional Hazards Regression Model. *Biometrika* 1982;69(1): 239-41
618. Gutierrez RG. Parametric frailty and shared frailty survival models
The Stata Journal 2002;2(1):22-44.
619. Thygesen LC, Ersboll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. *Eur J Epidemiol* 2014;29(8):551-8. doi: 10.1007/s10654-013-9873-0
620. Ali AM, Schmidt MK, Bolla MK, et al. Alcohol consumption and survival after a breast cancer diagnosis: a literature-based meta-analysis and collaborative analysis of data for 29,239 cases. *Cancer Epidemiol Biomarkers Prev* 2014;23(6):934-45. doi: 10.1158/1055-9965.EPI-13-0901 [published Online First: 2014/03/19]
621. Stott DJ. Alcohol and mortality in older people: understanding the J-shaped curve. *Age Ageing* 2020;49(3):332-33. doi: 10.1093/ageing/afaa027 [published Online First: 2020/04/29]
622. Browman GP, Mohide EA, Willan A, et al. Association between smoking during radiotherapy and prognosis in head and neck cancer: a follow-up study. *Head Neck* 2002;24(12):1031-7. doi: 10.1002/hed.10168
623. Hoff CM, Grau C, Overgaard J. Effect of smoking on oxygen delivery and outcome in patients treated with radiotherapy for head and neck squamous cell carcinoma--a prospective study. *Radiother Oncol* 2012;103(1):38-44. doi: 10.1016/j.radonc.2012.01.011 [published Online First: 2012/03/06]
624. Nordsmark M, Overgaard J. Tumor hypoxia is independent of hemoglobin and prognostic for loco-regional tumor control after primary radiotherapy in advanced

head and neck cancer. *Acta Oncol* 2004;43(4):396-403. [published Online First: 2004/08/12]
625. Brennan JA, Boyle JO, Koch WM, et al. Association between cigarette smoking and mutation of the p53 gene in squamous-cell carcinoma of the head and neck. *N Engl J Med* 1995;332(11):712-7. doi: 10.1056/NEJM199503163321104 [published Online First: 1995/03/16]
626. Yanbaeva DG, Dentener MA, Creutzberg EC, et al. Systemic effects of smoking. *Chest* 2007;131(5):1557-66. doi: 10.1378/chest.06-2179
627. Szabo G. Consequences of alcohol consumption on host defence. *Alcohol Alcohol* 1999;34(6):830-41. [published Online First: 2000/02/05]
628. Kreimer AR, Clifford GM, Snijders PJ, et al. HPV16 semiquantitative viral load and serologic biomarkers in oral and oropharyngeal squamous cell carcinomas. *Int J Cancer* 2005;115(2):329-32. doi: 10.1002/ijc.20872
629. Burd EM. Human Papillomavirus Laboratory Testing: the Changing Paradigm. *Clin Microbiol Rev* 2016;29(2):291-319. doi: 10.1128/CMR.00013-15 [published Online First: 2016/02/26]
630. Mell LK, Dignam JJ, Salama JK, et al. Predictors of competing mortality in advanced head and neck cancer. *J Clin Oncol* 2010;28(1):15-20. doi: 10.1200/JCO.2008.20.9288 [published Online First: 2009/11/26]
631. Smith Sehdev AE, Hutchins GM. Problems with proper completion and accuracy of the cause-of-death statement. *Arch Intern Med* 2001;161(2):277-84. [published Online First: 2001/02/15]
632. Corley J, Cox SR, Harris SE, et al. Epigenetic signatures of smoking associate with cognitive function, brain structure, and mental and physical health outcomes in the Lothian Birth Cohort 1936. *Transl Psychiatry* 2019;9(1):248. doi: 10.1038/s41398-019-0576-5 [published Online First: 2019/10/09]
633. McCartney DL, Hillary RF, Stevenson AJ, et al. Epigenetic prediction of complex traits and death. *Genome Biol* 2018;19(1):136. doi: 10.1186/s13059-018-1514-1 [published Online First: 2018/09/28]
634. Min J, Hemani G, Davey Smith G, et al. Meffil: efficient normalisation and analysis of very large DNA methylation samples. *bioRxiv* 2017 doi: 10.1101/125963
635. Zhang Y, Florath I, Saum KU, et al. Self-reported smoking, serum cotinine, and blood DNA methylation. *Environ Res* 2016;146:395-403. doi: 10.1016/j.envres.2016.01.026 [published Online First: 2016/02/02]
636. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;13:86. doi: 10.1186/1471-2105-13-86
637. Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. *J Stat Comput Simul* 2018;88(3):533-52. doi: 10.1080/00949655.2017.1397151 [published Online First: 2018/01/13]
638. Philibert R, Hollenbeck N, Andersen E, et al. Reversion of AHRR Demethylation Is a Quantitative Biomarker of Smoking Cessation. *Front Psychiatry* 2016;7:55. doi: 10.3389/fpsyt.2016.00055 [published Online First: 2016/04/20]
639. Ferlay J CM, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, Znaor A, Bray F. . Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer* 2018 (18.11.2018):In press. doi: https://doi.org/10.1002/ijc.31937
640. Nichols AC, Palma DA, Dhaliwal SS, et al. The epidemic of human papillomavirus and oropharyngeal cancer in a Canadian population. *Curr Oncol* 2013;20(4):212-9. doi: 10.3747/co.20.1375 [published Online First: 2013/08/02]
641. Mirghani H, Blanchard P. Treatment de-escalation for HPV-driven oropharyngeal cancer: Where do we stand? *Clin Transl Radiat Oncol* 2018;8:4-11. doi: 10.1016/j.ctro.2017.10.005 [published Online First: 2018/03/30]

642. Dugue PA, Bassett JK, Joo JE, et al. DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *Int J Cancer* 2018;142(8):1611-19. doi: 10.1002/ijc.31189 [published Online First: 2017/12/03]
643. Gillison ML, D'Souza G, Westra W, et al. Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. *J Natl Cancer I* 2008;100(6):407-20.
644. Schimansky S, Lang S, Beynon R, et al. Association between comorbidity and survival in head and neck cancer: Results from Head and Neck 5000. *Head Neck* 2019;41(4):1053-62. doi: 10.1002/hed.25543 [published Online First: 2018/12/15]
645. de Graeff A, de Leeuw JR, Ros WJ, et al. Sociodemographic factors and quality of life as prognostic indicators in head and neck cancer. *Eur J Cancer* 2001;37(3):332-9.
646. Hollander D, Kampman E, van Herpen CM. Pretreatment body mass index and head and neck cancer outcome: A review of the literature. *Crit Rev Oncol Hematol* 2015;96(2):328-38. doi: 10.1016/j.critrevonc.2015.06.002 [published Online First: 2015/06/29]
647. Hughes RA, Heron J, Sterne JAC, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International journal of epidemiology* 2019 doi: 10.1093/ije/dyz032 [published Online First: 2019/03/18]
648. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27(2):157-72; discussion 207-12. doi: 10.1002/sim.2929 [published Online First: 2007/06/15]
649. Royston P. Flexible parametric alternatives to the Cox model, and more. *Stata Journal* 2001;1:1-28.
650. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* 2002;21(15):2175-97. doi: 10.1002/sim.1203 [published Online First: 2002/09/05]
651. Ensor J, Riley RD, Jowett S, et al. Prediction of risk of recurrence of venous thromboembolism following treatment for a first unprovoked venous thromboembolism: systematic review, prognostic model and clinical decision rule, and economic evaluation. *Health Technol Assess* 2016;20(12):i-xxxiii, 1-190. doi: 10.3310/hta20120 [published Online First: 2016/02/18]
652. Lambert PC. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009;9:265–90.
653. Ng R, Kornas K, Sutradhar R, et al. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. *Diagn Progn Res* 2018;2:4. doi: 10.1186/s41512-018-0026-5 [published Online First: 2018/02/07]
654. Richard D. Riley vdW, Danielle, Peter Croft, Karel G.M. Moons Prognosis Research in Healthcare: Concepts, Methods, and Impact. first ed. Oxford, United Kingdom: Oxford University Press 2019.
655. Lambert PC. Sensitivity analysis to location of knots (proportional hazards) 2017 [08.10.19]. Available from: https://pclambert.net/software/stpm2/knot_positions_sensitivity/ accessed 08.10.2019.
656. Royston P SW. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Chichester: John Wiley & Sons, Ltd 2008.
657. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004;23(13):2109-23. doi: 10.1002/sim.1802 [published Online First: 2004/06/24]
658. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38(7):1276-96. doi: 10.1002/sim.7992 [published Online First: 2018/10/26]

659. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med* 2019;38(7):1262-75. doi: 10.1002/sim.7993 [published Online First: 2018/10/23]
660. Hudda MT, Fewtrell MS, Haroun D, et al. Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data. *BMJ* 2019;366:l4293. doi: 10.1136/bmj.l4293 [published Online First: 2019/07/26]
661. Levine ME, Hosgood HD, Chen B, et al. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging (Albany NY)* 2015;7(9):690-700. doi: 10.18632/aging.100809
662. Whittle R, Royle KL, Jordan KP, et al. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagn Progn Res* 2017;1:1. doi: 10.1186/s41512-016-0006-6 [published Online First: 2017/02/08]
663. Kubala MH, Punj V, Placencio-Hickok VR, et al. Plasminogen Activator Inhibitor-1 Promotes the Recruitment and Polarization of Macrophages in Cancer. *Cell Rep* 2018;25(8):2177-91 e7. doi: 10.1016/j.celrep.2018.10.082 [published Online First: 2018/11/22]
664. Duffy MJ, O'Donovan N, McDermott E, et al. Validated biomarkers: The key to precision treatment in patients with breast cancer. *Breast* 2016;29:192-201. doi: 10.1016/j.breast.2016.07.009 [published Online First: 2016/08/16]
665. Mengele K, Napieralski R, Magdolen V, et al. Characteristics of the level-of-evidence-1 disease forecast cancer biomarkers uPA and its inhibitor PAI-1. *Expert Rev Mol Diagn* 2010;10(7):947-62. doi: 10.1586/erm.10.73 [published Online First: 2010/10/23]
666. Schiegnitz E, Kammerer PW, Rode K, et al. Growth differentiation factor 15 as a radiation-induced marker in oral carcinoma increasing radiation resistance. *J Oral Pathol Med* 2016;45(1):63-9. doi: 10.1111/jop.12323 [published Online First: 2015/04/17]
667. Yang CZ, Ma J, Luo QQ, et al. Elevated level of serum growth differentiation factor 15 is associated with oral leukoplakia and oral squamous cell carcinoma. *J Oral Pathol Med* 2014;43(1):28-34. doi: 10.1111/jop.12091 [published Online First: 2013/05/29]
668. Zhang L, Yang X, Pan HY, et al. Expression of growth differentiation factor 15 is positively correlated with histopathological malignant grade and in vitro cell proliferation in oral squamous cell carcinoma. *Oral Oncol* 2009;45(7):627-32. doi: 10.1016/j.oraloncology.2008.07.017 [published Online First: 2008/09/23]
669. Tang X, Hu YJ, Ju WT, et al. Elevated growth differentiating factor 15 expression predicts long-term benefit of docetaxel, cisplatin and 5-fluorouracil induction chemotherapy in patients with oral cancer. *Oncol Lett* 2018;15(5):8118-24. doi: 10.3892/ol.2018.8324 [published Online First: 2018/05/08]
670. Yang CZ, Ma J, Zhu DW, et al. GDF15 is a potential predictive biomarker for TPF induction chemotherapy and promotes tumorigenesis and progression in oral squamous cell carcinoma. *Ann Oncol* 2014;25(6):1215-22. doi: 10.1093/annonc/mdu120 [published Online First: 2014/03/29]
671. Langdon R, Richmond R, Elliott HR, et al. Identifying epigenetic biomarkers of established prognostic factors and survival in a clinical cohort of individuals with oropharyngeal cancer. *BioRxiv* 2019 doi: http://dx.doi.org/10.1101/679316.
672. Royston P. Multiple imputation of missing values. *Stata Journal* 2004;4:227–41.
673. McEwen LM, Jones MJ, Lin DTS, et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clin Epigenetics* 2018;10(1):123. doi: 10.1186/s13148-018-0556-2 [published Online First: 2018/10/18]
674. Kaushik AK, DeBerardinis RJ. Applications of metabolomics to study cancer metabolism. *Biochim Biophys Acta Rev Cancer* 2018;1870(1):2-14. doi: 10.1016/j.bbcan.2018.04.009 [published Online First: 2018/04/28]

675. Beger RD. A review of applications of metabolomics in cancer. *Metabolites* 2013;3(3):552-74. doi: 10.3390/metabo3030552
676. Spratlin JL, Serkova NJ, Eckhardt SG. Clinical applications of metabolomics in oncology: a review. *Clin Cancer Res* 2009;15(2):431-40. doi: 10.1158/1078-0432.CCR-08-1059
677. Mansara PP, Deshpande RA, Vaidya MM, et al. Differential Ratios of Omega Fatty Acids (AA/EPA+DHA) Modulate Growth, Lipid Peroxidation and Expression of Tumor Regulatory MARBPs in Breast Cancer Cell Lines MCF7 and MDA-MB-231. *PLoS One* 2015;10(9):e0136542. doi: 10.1371/journal.pone.0136542 [published Online First: 2015/09/02]
678. Apte SA, Cavazos DA, Whelan KA, et al. A low dietary ratio of omega-6 to omega-3 Fatty acids may delay progression of prostate cancer. *Nutr Cancer* 2013;65(4):556-62. doi: 10.1080/01635581.2013.775316 [published Online First: 2013/05/11]
679. Health N. Blood Biomarker Analysis, 2020 [Available from: nightingalehealth.com/research/blood-biomarker-analysis accessed 29.06.20.
680. Santos Ferreira DL, Maple HJ, Goodwin M, et al. The Effect of Pre-Analytical Conditions on Blood Metabolomics in Epidemiological Studies. *Metabolites* 2019;9(4) doi: 10.3390/metabo9040064 [published Online First: 2019/04/17]
681. Playdon MC, Joshi AD, Tabung FK, et al. Metabolomics Analytics Workflow for Epidemiological Research: Perspectives from the Consortium of Metabolomics Studies (COMETS). *Metabolites* 2019;9(7) doi: 10.3390/metabo9070145 [published Online First: 2019/07/20]
682. Dudoit SPS, J.; Boldrick, J.C. . Multiple Hypothesis Testing in Microarray Experiments. *Stat Sci* 2003(18):71–103.
683. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 2016;374(2065):20150202. doi: 10.1098/rsta.2015.0202 [published Online First: 2016/03/10]
684. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* 2008;32(4):361-9. doi: 10.1002/gepi.20310 [published Online First: 2008/02/14]
685. Braeken J, van Assen M. An empirical Kaiser criterion. *Psychol Methods* 2017;22(3):450-66. doi: 10.1037/met0000074 [published Online First: 2016/04/01]
686. Statistics How To. 2020 [16.01.20]. Available from: https://www.statisticshowto.datasciencecentral.com/varimax-rotation-definition/ accessed 16.01.20.
687. Lee KJ, Carlin JB. Multiple imputation in the presence of non-normal data. *Stat Med* 2017;36(4):606-17. doi: 10.1002/sim.7173 [published Online First: 2016/11/20]
688. Hippel PTv. Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociological Methods & Research* 2012;45(1):105-38.
689. Aretz I, Meierhofer D. Advantages and Pitfalls of Mass Spectrometry Based Metabolome Profiling in Systems Biology. *Int J Mol Sci* 2016;17(5) doi: 10.3390/ijms17050632 [published Online First: 2016/04/30]
690. Gromski PS, Xu Y, Kotze HL, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites* 2014;4(2):433-52. doi: 10.3390/metabo4020433 [published Online First: 2014/06/25]
691. Pollet TV, van der Meij, L. To Remove or not to Remove: the Impact of Outlier Handling on Significance Testing in Testosterone Data. *Adaptive Human Behavior and Physiology* 2017(3):43–60 doi: 10.1007/s40750-016-0050-z
692. Statistics How To. Empirical Rule: What is it? [cited 20.01.20. Available from: https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/ accessed 20.01.20.
693. Do KT, Wahl S, Raffler J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies.

Metabolomics 2018;14(10):128. doi: 10.1007/s11306-018-1420-2 [published Online First: 2019/03/05]
694. Bhattacharyya S, Ahmed AT, Arnold M, et al. Metabolomic signature of exposure and response to citalopram/escitalopram in depressed outpatients. *Transl Psychiatry* 2019;9(1):173. doi: 10.1038/s41398-019-0507-5 [published Online First: 2019/07/06]
695. Long T, Hicks M, Yu HC, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet* 2017;49(4):568-78. doi: 10.1038/ng.3809 [published Online First: 2017/03/07]
696. Leys C, Ley C, Klein O, et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*;49(4):764-66.
697. Ronacher K, Chegou NN, Kleynhans L, et al. Distinct serum biosignatures are associated with different tuberculosis treatment outcomes. *Tuberculosis (Edinb)* 2019;118:101859. doi: 10.1016/j.tube.2019.101859 [published Online First: 2019/08/23]
698. Yu B, de Vries PS, Metcalf GA, et al. Whole genome sequence analysis of serum amino acid levels. *Genome Biol* 2016;17(1):237. doi: 10.1186/s13059-016-1106-x [published Online First: 2016/11/26]
699. Madeira C, Alheira FV, Calcia MA, et al. Blood Levels of Glutamate and Glutamine in Recent Onset and Chronic Schizophrenia. *Front Psychiatry* 2018;9:713. doi: 10.3389/fpsyt.2018.00713 [published Online First: 2019/01/09]
700. Schug ZT, Vande Voorde J, Gottlieb E. The metabolic fate of acetate in cancer. *Nat Rev Cancer* 2016;16(11):708-17. doi: 10.1038/nrc.2016.87 [published Online First: 2016/10/25]
701. Cheng C, Geng F, Cheng X, et al. Lipid metabolism reprogramming and its potential targets in cancer. *Cancer Commun (Lond)* 2018;38(1):27. doi: 10.1186/s40880-018-0301-4 [published Online First: 2018/05/23]
702. Hosios AM, Vander Heiden MG. Acetate metabolism in cancer cells. *Cancer Metab* 2014;2(1):27. doi: 10.1186/s40170-014-0027-y [published Online First: 2014/12/17]
703. Kamphorst JJ, Chung MK, Fan J, et al. Quantitative analysis of acetyl-CoA production in hypoxic cancer cells reveals substantial contribution from acetate. *Cancer Metab* 2014;2:23. doi: 10.1186/2049-3002-2-23 [published Online First: 2015/02/12]
704. Pietrocola F, Galluzzi L, Bravo-San Pedro JM, et al. Acetyl coenzyme A: a central metabolite and second messenger. *Cell Metab* 2015;21(6):805-21. doi: 10.1016/j.cmet.2015.05.014 [published Online First: 2015/06/04]
705. Lyssiotis CA, Cantley LC. Acetate fuels the cancer engine. *Cell* 2014;159(7):1492-4. doi: 10.1016/j.cell.2014.12.009 [published Online First: 2014/12/20]
706. Gao X, Lin SH, Ren F, et al. Acetate functions as an epigenetic metabolite to promote lipid synthesis under hypoxia. *Nat Commun* 2016;7:11960. doi: 10.1038/ncomms11960 [published Online First: 2016/07/01]
707. Kuo CY, Ann DK. When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance. *Cancer Commun (Lond)* 2018;38(1):47. doi: 10.1186/s40880-018-0317-9 [published Online First: 2018/07/13]
708. (NICE) TNIfHaCE. Alcohol use disorders. Diagnosis, assessment and management of harmful drinking and alcohol dependence. 2010 [12.11.19]. Available from: https://www.nice.org.uk/guidance/cg115/documents/alcohol-dependence-and-harmful-alcohol-use-full-guideline2 accessed 12.11.19.
709. Ostermann M, Kashani K, Forni LG. The two sides of creatinine: both as bad as each other? *J Thorac Dis* 2016;8(7):E628-30. doi: 10.21037/jtd.2016.05.36 [published Online First: 2016/08/09]
710. Thongprayoon C, Cheungpasitporn W, Kashani K. Serum creatinine level, a surrogate of muscle mass, predicts mortality in critically ill patients. *J Thorac Dis* 2016;8(5):E305-11. doi: 10.21037/jtd.2016.03.62 [published Online First: 2016/05/11]
711. Pin F, Couch ME, Bonetto A. Preservation of muscle mass as a strategy to reduce the toxic effects of cancer chemotherapy on body composition. *Curr Opin Support Palliat*

Care 2018;12(4):420-26. doi: 10.1097/SPC.0000000000000382 [published Online First: 2018/08/21]
712. Davis MP, Panikkar R. Sarcopenia associated with chemotherapy and targeted agents for cancer therapy. *Ann Palliat Med* 2019;8(1):86-101. doi: 10.21037/apm.2018.08.02 [published Online First: 2018/12/12]
713. Chargi N, Bril SI, Emmelot-Vonk MH, et al. Sarcopenia is a prognostic factor for overall survival in elderly patients with head-and-neck cancer. *Eur Arch Otorhinolaryngol* 2019;276(5):1475-86. doi: 10.1007/s00405-019-05361-4 [published Online First: 2019/03/05]
714. Baxi SS, Schwitzer E, Jones LW. A review of weight loss and sarcopenia in patients with head and neck cancer treated with chemoradiation. *Cancers Head Neck* 2016;1:9. doi: 10.1186/s41199-016-0010-0 [published Online First: 2016/08/17]
715. Ganju RG, Morse R, Hoover A, et al. The impact of sarcopenia on tolerance of radiation and outcome in patients with head and neck cancer receiving chemoradiation. *Radiother Oncol* 2019;137:117-24. doi: 10.1016/j.radonc.2019.04.023 [published Online First: 2019/05/16]
716. Thongprayoon C, Cheungpasitporn W, Kittanamongkolchai W, et al. Prognostic Importance of Low Admission Serum Creatinine Concentration for Mortality in Hospitalized Patients. *Am J Med* 2017;130(5):545-54 e1. doi: 10.1016/j.amjmed.2016.11.020 [published Online First: 2016/12/22]
717. Howes N, Atkinson C, Thomas S, et al. Immunonutrition for patients undergoing surgery for head and neck cancer. *Cochrane Database Syst Rev* 2018;8:CD010954. doi: 10.1002/14651858.CD010954.pub2 [published Online First: 2018/08/31]
718. de Luis DA, Izaola O, Cuellar L, et al. A randomized clinical trial with two doses of a omega 3 fatty acids oral and arginine enhanced formula in clinical and biochemical parameters of head and neck cancer ambulatory patients. *Eur Rev Med Pharmacol Sci* 2013;17(8):1090-4. [published Online First: 2013/05/11]
719. de Luis DA, Izaola O, Aller R, et al. A randomized clinical trial with two omega 3 fatty acid enhanced oral supplements in head and neck cancer ambulatory patients. *Eur Rev Med Pharmacol Sci* 2008;12(3):177-81. [published Online First: 2008/08/15]
720. Freitas RDS, Campos MM. Protective Effects of Omega-3 Fatty Acids in Cancer-Related Complications. *Nutrients* 2019;11(5) doi: 10.3390/nu11050945 [published Online First: 2019/05/01]
721. Stableforth WD, Thomas S, Lewis SJ. A systematic review of the role of immunonutrition in patients undergoing surgery for head and neck cancer. *Int J Oral Maxillofac Surg* 2009;38(2):103-10. doi: 10.1016/j.ijom.2008.12.008 [published Online First: 2009/01/16]
722. Tzoulaki I, Ebbels TM, Valdes A, et al. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am J Epidemiol* 2014;180(2):129-39. doi: 10.1093/aje/kwu143 [published Online First: 2014/06/27]
723. Vinaixa M, Samino S, Saez I, et al. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites* 2012;2(4):775-95. doi: 10.3390/metabo2040775 [published Online First: 2012/01/01]
724. Markley JL, Bruschweiler R, Edison AS, et al. The future of NMR-based metabolomics. *Curr Opin Biotechnol* 2017;43:34-40. doi: 10.1016/j.copbio.2016.08.001 [published Online First: 2016/09/01]
725. Nightingale Health 2019 [12.11.19]. Available from: https://nightingalehealth.com/ accessed 12.11.19.
726. Hernandes VV, Barbas C, Dudzik D. A review of blood sample handling and pre-processing for metabolomics studies. *Electrophoresis* 2017;38(18):2232-41. doi: 10.1002/elps.201700086 [published Online First: 2017/05/26]
727. Imhoi Koo, Xiaoli Wei, Zhang X. Chapter Sixteen - Analysis of Metabolomic Profiling Data Acquired on GC–MS. *Methods in Enzymology*;543:315-24.

728. Metabolon. The Five Key Elements of a Successful Metabolomics Study [23.01.2020]. Available from: https://www.metabolon.com/application/files/2415/0388/1050/The-5-Key-Elements-of-a-Successful-Metabolomics-Study.pdf accessed 23.01.2020.
729. Biobank. Biomarkers [23.01.2020]. Available from: https://www.ukbiobank.ac.uk/uk-biobank-biomarker-panel/ accessed 23.01.2020.
730. Dunn WB, Broadhurst DI, Atherton HJ, et al. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* 2011;40(1):387-426. doi: 10.1039/b906712b [published Online First: 2010/08/19]
731. Euser AM, Zoccali C, Jager KJ, et al. Cohort studies: prospective versus retrospective. *Nephron Clin Pract* 2009;113(3):c214-7. doi: 10.1159/000235241 [published Online First: 2009/08/20]
732. Song JW, Chung KC. Observational studies: cohort and case-control studies. *Plast Reconstr Surg* 2010;126(6):2234-42. doi: 10.1097/PRS.0b013e3181f44abc [published Online First: 2010/08/11]
733. Howe CJ, Cole SR, Lau B, et al. Selection Bias Due to Loss to Follow Up in Cohort Studies. *Epidemiology* 2016;27(1):91-7. doi: 10.1097/EDE.0000000000000409 [published Online First: 2015/10/21]
734. Zhang Y, Hedo R, Rivera A, et al. Post hoc power analysis: is it an informative and meaningful analysis? *Gen Psychiatr* 2019;32(4):e100069. doi: 10.1136/gpsych-2019-100069 [published Online First: 2019/09/26]
735. Kadam P, Bhalerao S. Sample size calculation. *Int J Ayurveda Res* 2010;1(1):55-7. doi: 10.4103/0974-7788.59946 [published Online First: 2010/06/10]
736. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? *Indian J Psychol Med* 2013;35(2):121-6. doi: 10.4103/0253-7176.116232 [published Online First: 2013/09/21]
737. Levine M, Ensom MH. Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy* 2001;21(4):405-9. doi: 10.1592/phco.21.5.405.34503 [published Online First: 2001/04/20]
738. Hoenig J, Heisey, DM. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician* 2001;55
739. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 1994;121(3):200-6. doi: 10.7326/0003-4819-121-3-199408010-00008 [published Online First: 1994/08/01]
740. Lenth RV. Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician* 2001;55:187-93.
741. Plate JDJ, Borggreve AS, van Hillegersberg R, et al. Post Hoc Power Calculation: Observing the Expected. *Ann Surg* 2019;269(1):e11. doi: 10.1097/SLA.0000000000002910 [published Online First: 2018/07/07]
742. Wissing MD, Greenwald ZR, Franco EL. Improving the reporting of cancer-specific mortality and survival in research using cancer registry data. *Cancer Epidemiol* 2019;59:232-35. doi: 10.1016/j.canep.2019.02.004 [published Online First: 2019/03/06]
743. Simpson MC, Massa ST, Boakye EA, et al. Primary Cancer vs Competing Causes of Death in Survivors of Head and Neck Cancer. *JAMA Oncol* 2018;4(2):257-59. doi: 10.1001/jamaoncol.2017.4478 [published Online First: 2017/12/30]
744. Liberti MV, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem Sci* 2016;41(3):211-18. doi: 10.1016/j.tibs.2015.12.001 [published Online First: 2016/01/19]
745. Burgess S, Timpson NJ, Ebrahim S, et al. Mendelian randomization: where are we now and where are we going? *International journal of epidemiology* 2015;44(2):379-88. doi: 10.1093/ije/dyv108 [published Online First: 2015/06/19]

746. Lawlor DA, Harbord RM, Sterne JA, et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008;27(8):1133-63. doi: 10.1002/sim.3034 [published Online First: 2007/09/22]
747. Lawlor DA. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International journal of epidemiology* 2016;45(3):908-15. doi: 10.1093/ije/dyw127 [published Online First: 2016/07/19]
748. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 2018;7 doi: 10.7554/eLife.34408 [published Online First: 2018/05/31]
749. Walker VM, Davies NM, Hemani G, et al. Using the MR-Base platform to investigate risk factors and drug targets for thousands of phenotypes. *Wellcome Open Res* 2019;4:113. doi: 10.12688/wellcomeopenres.15334.2 [published Online First: 2019/11/20]
750. Gieger C, Geistlinger L, Altmaier E, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* 2008;4(11):e1000282. doi: 10.1371/journal.pgen.1000282 [published Online First: 2008/12/02]
751. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003;14(3):300-6. [published Online First: 2003/07/16]
752. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *International journal of epidemiology* 2010;39(2):417-20. doi: 10.1093/ije/dyp334 [published Online First: 2009/11/21]
753. Lim Y, Wan Y, Vagenas D, et al. Salivary DNA methylation panel to diagnose HPV-positive and HPV-negative head and neck cancers. *BMC Cancer* 2016;16(1):749. doi: 10.1186/s12885-016-2785-0 [published Online First: 2016/09/25]
754. Yoshizawa JM, Schafer CA, Schafer JJ, et al. Salivary biomarkers: toward future clinical and diagnostic utilities. *Clin Microbiol Rev* 2013;26(4):781-91. doi: 10.1128/CMR.00021-13 [published Online First: 2013/10/05]
755. Lee KD, Lee HS, Jeon CH. Body fluid biomarkers for early detection of head and neck squamous cell carcinomas. *Anticancer Res* 2011;31(4):1161-7.
756. Roi A, Rusu LC, Roi CI, et al. A New Approach for the Diagnosis of Systemic and Oral Diseases Based on Salivary Biomolecules. *Dis Markers* 2019;2019:8761860. doi: 10.1155/2019/8761860 [published Online First: 2019/02/17]
757. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60(3):431-49. doi: 10.1002/bimj.201700067 [published Online First: 2018/01/02]
758. Bujak R, Daghir-Wojtkowiak E, Kaliszan R, et al. PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-targeted Metabolomics Data. *Front Mol Biosci* 2016;3:35. doi: 10.3389/fmolb.2016.00035 [published Online First: 2016/08/11]
759. Kirpich A, Ainsworth EA, Wedow JM, et al. Variable selection in omics data: A practical evaluation of small sample sizes. *PLoS One* 2018;13(6):e0197910. doi: 10.1371/journal.pone.0197910 [published Online First: 2018/06/22]